Introduction to Network Analysis 030824 James Ward

Final Project

This project will require you to use all the techniques we have covered in class so far. Your goal is to produce a visualization or two, plus an accompanying 2-3 page analysis and description of the Enron email network. Enron was a U.S. energy trading company that became embroiled in a major accounting and financial scandal in 2001. The scandal and its aftermath caused the failure of Enron and the dissolution of its accounting firm, Arthur Anderson.

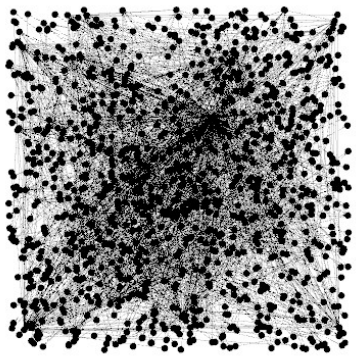
In your visualization(s) you can filter the network, but not too much. You might also use filters to help your analysis (e.g. filter out people of low centrality so you can focus on learning more about the people with high centrality) but not show those visualizations.

In your analysis, the following concepts should be discussed:

- Centrality
- Tie strength (e.g. how do the number of messages exchanged relate to tie strength - use the measures we discussed in class to quantify this)
- Clusters, cliques, communities
- Density
- Egocentric networks

Your analysis should go beyond simply reporting network metrics and describing the network structure. Who are the central people? What is their role in the network and the organization? (e.g. what kind of messages are they sending? to whom?) What are the big clusters and what does each represent? How did you find that out? What kinds of relationships exist? Are there strong personal relationships? Can you identify those by searching for important indicators in the email? What are the most important/strongest relationships? Are those determined by the number of interactions? By the strength shown in the content?

I begin by importing the Enron data into Gephi. The output is a basic cube of a network that I will begin to analyze.

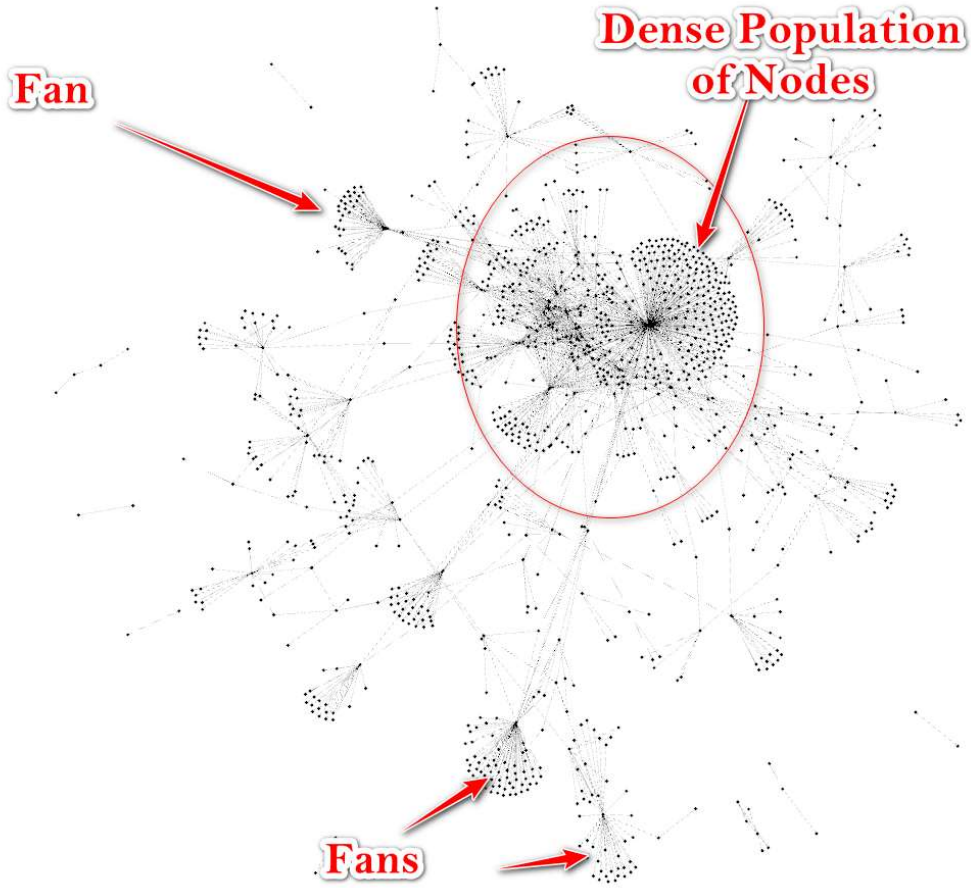


I will run a layout to begin to visualize the nodes and edges of this network. I like the Yifan Hu algorithm for visualization because it separates the nodes in an excellent way so that the connections can be easily seen.

In the layout, the network shows a dense population of nodes with individual clusters of nodes connected to single nodes within the dense population.

Total nodes: 1100

Edges: 1801

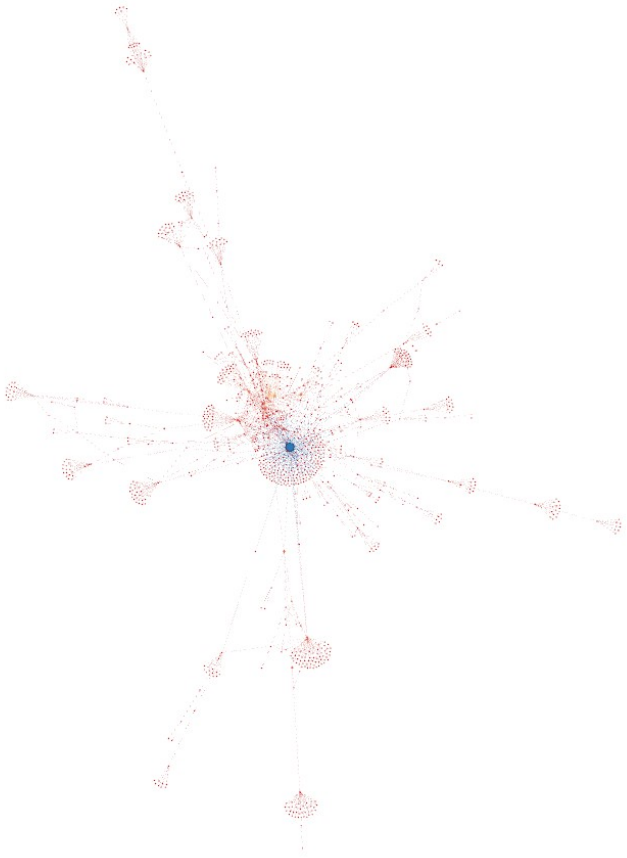


Betweenness Centrality

Initially, I am going to explore the Network Diameter of the network. The diameter can give an indication of the "spread" of the network and is a measure of how widely the nodes in the network are dispersed. A smaller diameter suggests that the network is more closely knit, where everyone is separated by only a few degrees. A larger diameter indicates that the network is more spread out, with some pairs of nodes being separated by many connections.

1. Betweenness Centrality: Betweenness centrality is a measure of centrality in a graph based on shortest paths. For every pair of vertices in a connected graph, there are one or more shortest paths between them. The betweenness centrality for each vertex is the number of these shortest paths that pass through the vertex.
2. Here's a more detailed breakdown:

3. **Shortest Paths**: The algorithm considers all shortest paths between pairs of nodes in the graph.
4. **Path Inclusion**: It calculates the fraction of these paths that pass through a given node (other than the start and end nodes).
5. **Centrality Score**: The betweenness centrality score for a node is the sum of these fractions over all pairs of nodes.



After running a color addition and to the nodes and adjusting the nodes by size with a minimum and maximum of 10 and 50, one can see that the fans are spread widely around the dense population of nodes in the center. The average path length is 3.814. Most of the fan clusters are tied to single nodes. I will observe these nodes shortly but first I want to look at the whole network.

When a node has a betweenness centrality of zero, it means that no shortest paths between any pair of nodes in the network pass through that node. Here are some implications and scenarios for a zero betweenness centrality:

- The node might be a peripheral node with only one connection, meaning it is at the edge of the network and does not act as a bridge or connector for other nodes.

- The node could be part of a clique or a tightly-knit group where all members are directly connected to each other, so there's no need for a path to go through that node to reach others within the same group.
- The node might exist in a disconnected component of the network that is not on any shortest path of the larger network.

On the other hand, a high betweenness centrality indicates that the node acts as a significant connector or bridge within the network. This means:

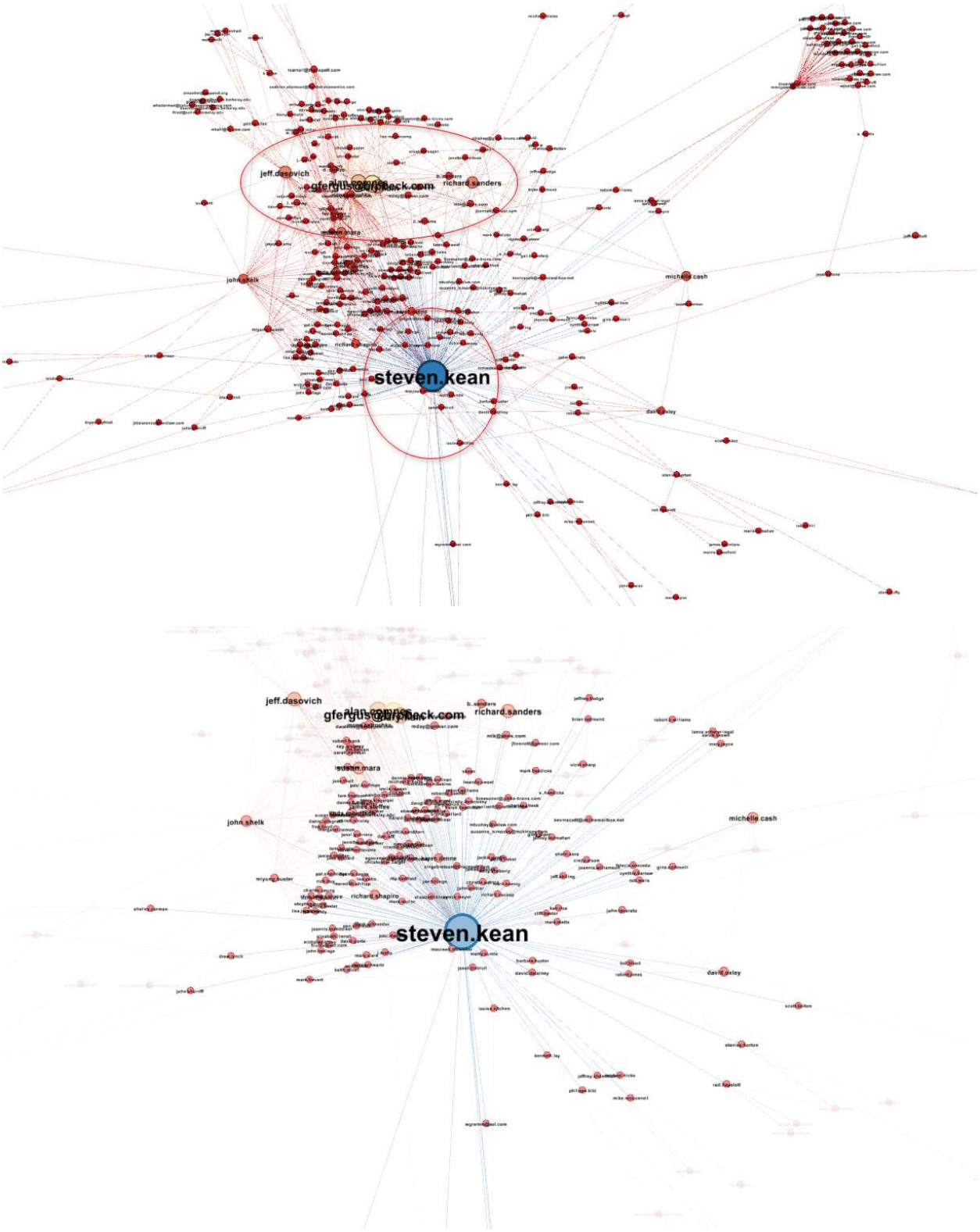
- The node is likely part of many shortest paths between various other nodes, suggesting it has a strategic position for controlling or influencing flow across the network.
- It might be a point of vulnerability; if this node were removed, it could cause increased path lengths or even disconnect parts of the network.
- In social networks, a person with high betweenness centrality could be someone who connects different social circles or has significant influence over information flow.

Overall, betweenness centrality highlights the potential for a node to exert control over interactions within the network by virtue of its position along the shortest paths.

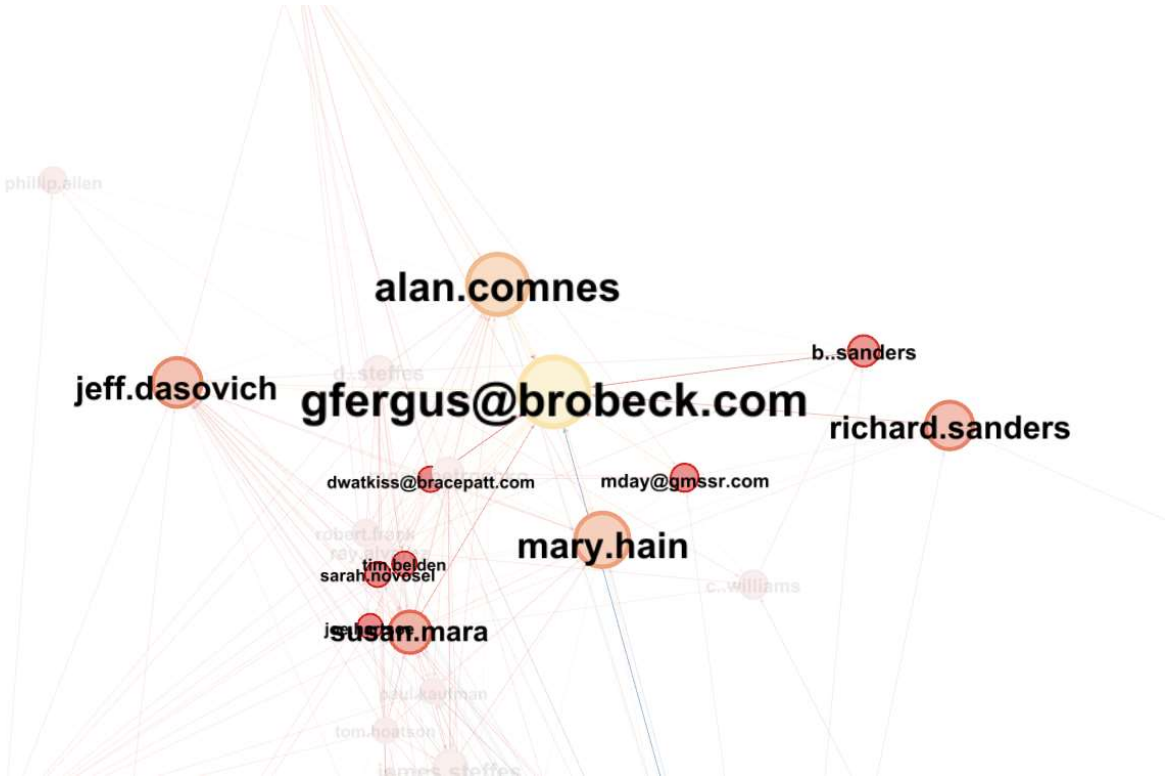
So, the analysis strategy is to analyze the nodes of higher betweenness centrality. See the list below.

Id	Betweenness Centrality
steven.kean	24195.428023
gfergus@brobeck.com	10778.421804
alan.comnes	8603.19738
mary.hain	7078.908822
jeff.dasovich	5861.445907
richard.sanders	5568.179038
vince.kaminski	4324.333333
j.kaminski	4235.666667
susan.mara	4215.697207
wolak@zia.stanford.edu	4003.666667
john.shelk	3578.10188
michelle.cash	3486.5
richard.shapiro	2502.266667
linda.robertson	2334.056862
james.steffes	2193.303229
david.oxley	1954.0
karen.denne	1714.396293
sally.beck	1600.0
beth.apollo	1560.0
b..sanders	1537.73373
d..steffes	1489.611829
mona.petrochko	1360.680664
m..presto	1341.0
douglass@arterhadden.com	1101.0
dana.davis	1085.0
thane.twiggs	1047.0
miyung.buster	956.519841
rcarroll@bracepatt.com	924.003501
ray.alvarez	868.461163
mlk@pkns.com	838.980159
c..williams	824.098413
mday@gmssr.com	763.748268
l..nicolay	749.2
jeff.bartlett	594.0
peggy.mahoney	537.715521
jeffrey.gossett	528.0
shirley.crenshaw	392.0
susan.scott	385.0
robert.frank	384.319729
phillip.allen	370.145238
rod.hayslett	321.0
david.delainey	319.169952
john.lavorato	259.0
kevinscott@onlinemailbox.net	257.083333

The following graph shows the degree of 2 which removes most of the nodes in the fan like structures. There are still nodes that fall outside the center with minimal connections and the more visible nodes with higher BC.

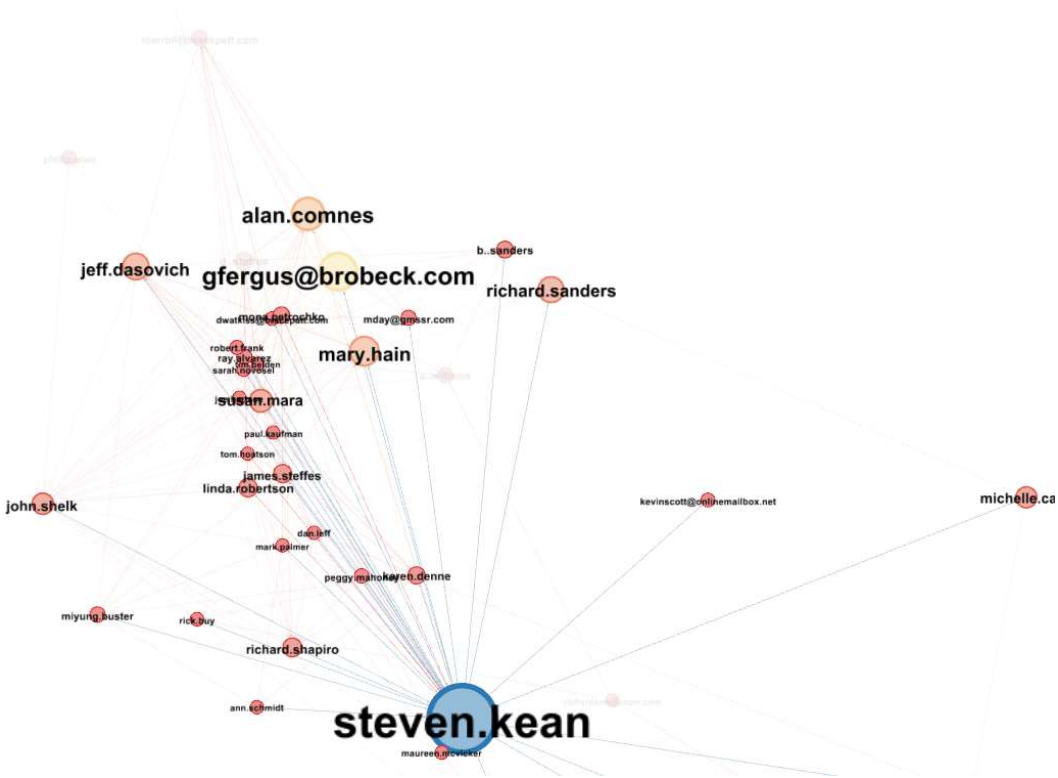


Steven.kean has a lot of one edge nodes. He does not end up in a significant modularity class but is nonetheless central to the company's operation.



After steven.kean, the nodes with high Betweeness Centrality are clustered around gferjus@brobeck.com, degree range = 10.

Steven.keen appears to be a geopolitical strategist for the company. His position would explain many of the single node contacts. He uses information to benefit the decisions made by the company. In addition to the single nodes, steven.kean connects with the other large BC nodes.



Closeness Centrality

Closeness centrality is a measure of how close a node is to all other nodes in a network. It is defined as the reciprocal of the sum of the shortest path distances from a given node to all other nodes in the graph. Thus, the more central a node is, the lower the total distance from it to all other nodes.

A high closeness centrality indicates that a node has shorter distances to all other nodes, meaning it can spread information very efficiently across the network. In other words, the node is central in the sense that it is relatively close to all other nodes.

Conversely, a closeness centrality of zero implies that a node is infinitely distant from the rest of the network, which can only happen in the case of a disconnected graph where there are nodes that cannot be reached from the node in question. In a connected graph, closeness centrality can never be zero because there is always some finite path to every other node.

Here's what different levels of closeness centrality signify:

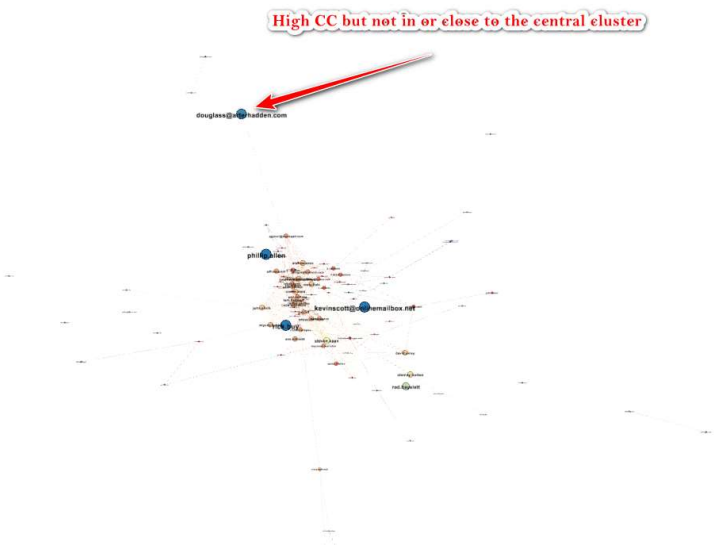
- **High Closeness Centrality**: The node is well-positioned within the network, allowing it to quickly interact with all other nodes. This might represent a person

who is active in a social network, a well-connected airport in a transportation network, or a key page in a web hyperlink network.

- **Low Closeness Centrality**: The node is on the periphery of the network and would take longer to spread information to all other nodes. This might represent a person with fewer social connections or an airport that offers fewer direct flights, thus requiring more connections to reach all other airports in the network.
- **Zero Closeness Centrality**: As mentioned, this would typically mean the node is in its own disconnected component of the network and cannot reach any other nodes, which is a theoretical extreme scenario in the case of non-connected graphs. In practical terms, this could represent an isolated user in a social network or a remote location in a transportation network.

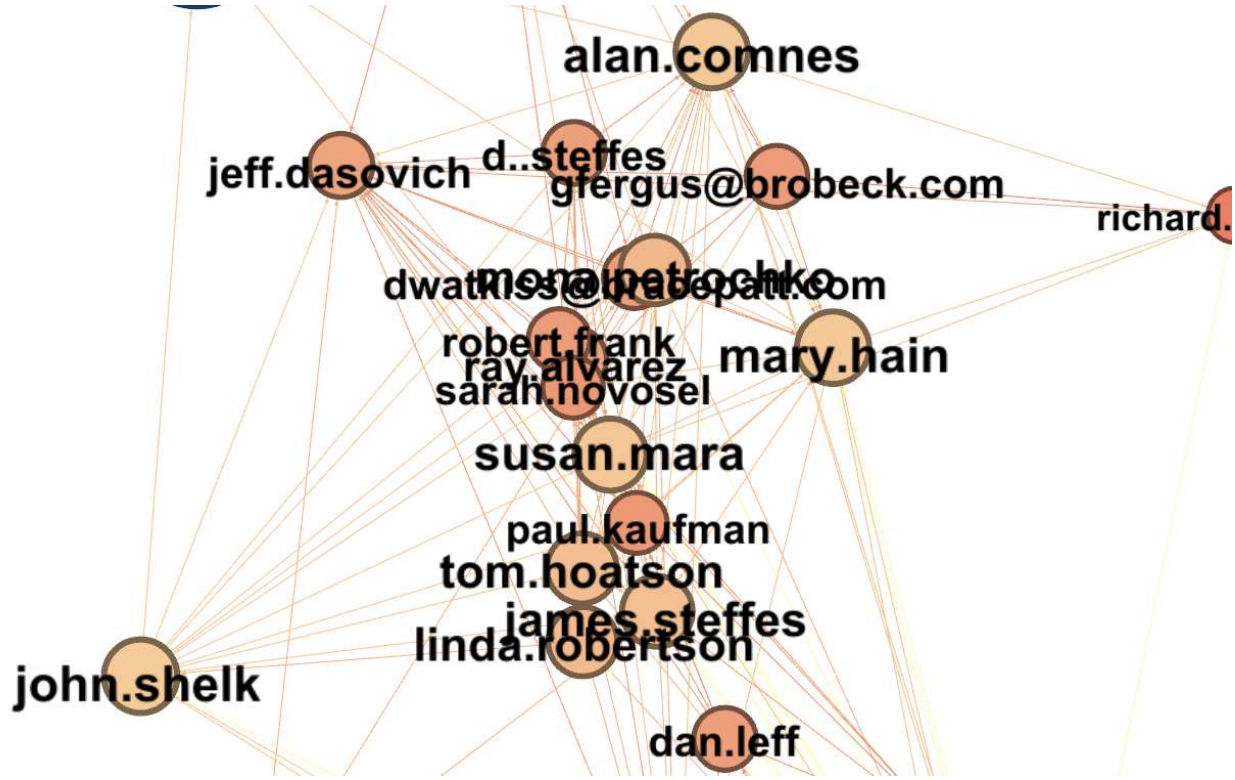
Id	Closeness Centrality
droark@velaw.com	1.0
mike.mcconnell	1.0
donna.scott	1.0
mark.haedicke	1.0
ruth.concannon	1.0
rick.buy	1.0
jeffrey.gossett	1.0
legalonline-compliance	1.0
phillip.allen	1.0
kevinscott@onlinemailbox.net	1.0
kenton.erwin	1.0
steven@iepa.com	1.0
m..tholt	1.0
stephanie.miller	1.0
sarah-joy.hunter	1.0
matt.harris	1.0
thane.twiggs	1.0
terri_l_ponce_de_leon@calpx....	1.0
barton.clark	1.0
paul.simons	1.0
sandi_j._thompson@calpx.com	1.0
jordan.mintz	1.0
linda.wehring	0.954545
douglass@arterhadden.com	0.954545
j.kaminski	0.95
mmcgown@velaw.com	0.947368
susan.scott	0.875
drew.fossum	0.875
marcus.nettelton	0.857143
susan.lopez	0.823529
barbara.gray	0.761905
beth.apollo	0.757576
jeffrey.hodge	0.75
jmball@ns.net	0.688525
k..allen	0.6875
vicki.sharp	0.666667
mariella.mahan	0.625
rod.hayslett	0.625
jeff.skilling	0.538462
j..kean	0.533333
steven.kean	0.506406
wolak@zia.stanford.edu	0.5
stanley.horton	0.434783
bwoertz@caiso.com	0.426415

Closeness Centrality measures proximity of nodes to central locations within the network. Some nodes which may not have a high proximity to the central cluster have a high (1.0) CC.



A high closeness centrality does not indicate importance all the time. In this network, nodes with high betweenness centrality have been different than the nodes with high Closeness Centrality.

There is a cluster within the 1.0 nodes that does include many of the nodes that were a part of the Betweenness Centrality nodes of importance.



Modularity Class

Modularity Class, often associated with the term "Modularity" in the context of network analysis, is a measure used to detect the presence of community structure in networks. Modularity quantifies the strength of division of a network into modules (also called groups, clusters, or communities).

Here's what modularity represents:

- **Modularity**: It is a scale value between -1 and 1 that measures the density of links inside communities compared to links between communities. In the calculation of modularity, networks are divided into groups, and the number of edges that fall within groups are compared to what would be expected on the basis of chance.
- **Modularity Class**: A modularity class is typically an output of algorithms that aim to detect community structure, assigning each node to a community or class. Nodes with the same modularity class label are considered part of the same community.

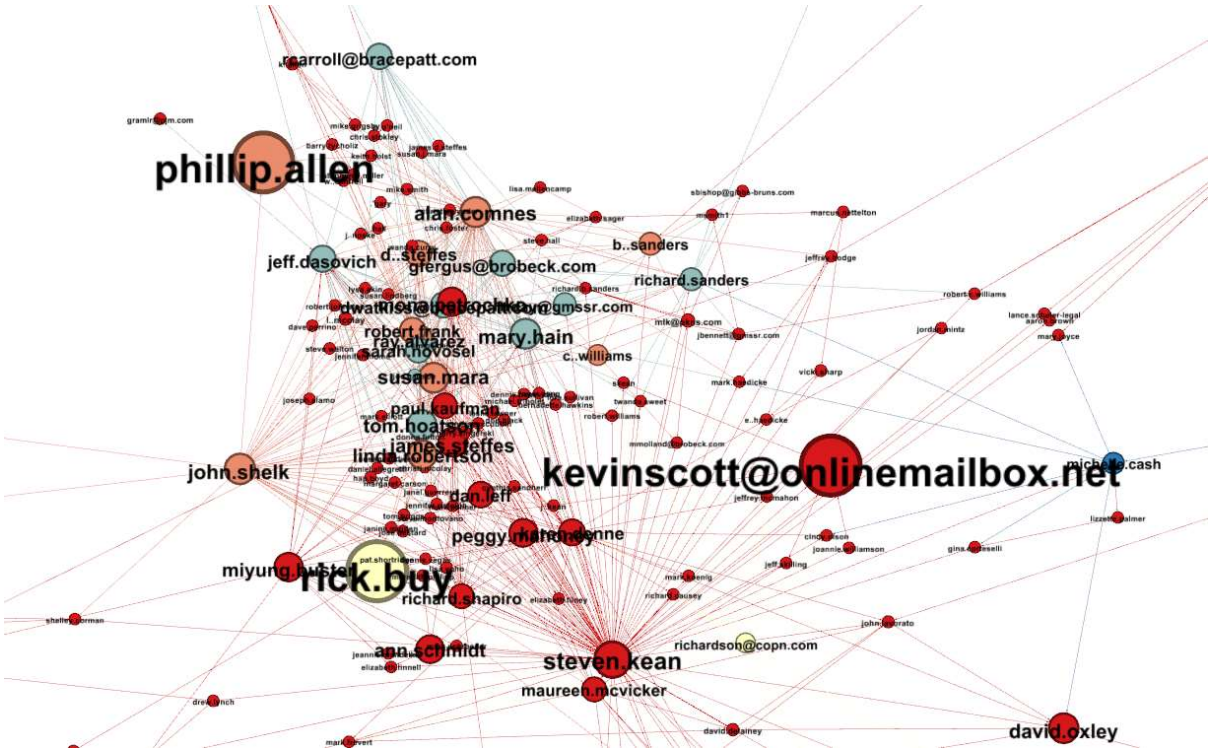
Regarding the values:

- **Modularity Class of Zero**: If all nodes have a modularity class of zero, it would suggest that the network does not exhibit any strong community structure as

defined by the algorithm used for modularity detection. It could be an indication that all nodes are considered part of a single community or that the community detection algorithm failed to find meaningful clusters in the data.

- **High Modularity Class**: A high modularity value would indicate that the network's structure is characterized by dense connections within communities and sparse connections between them, implying a strong community structure. Nodes with the same high modularity class value are grouped together, meaning they have more connections amongst themselves than with nodes of other classes.

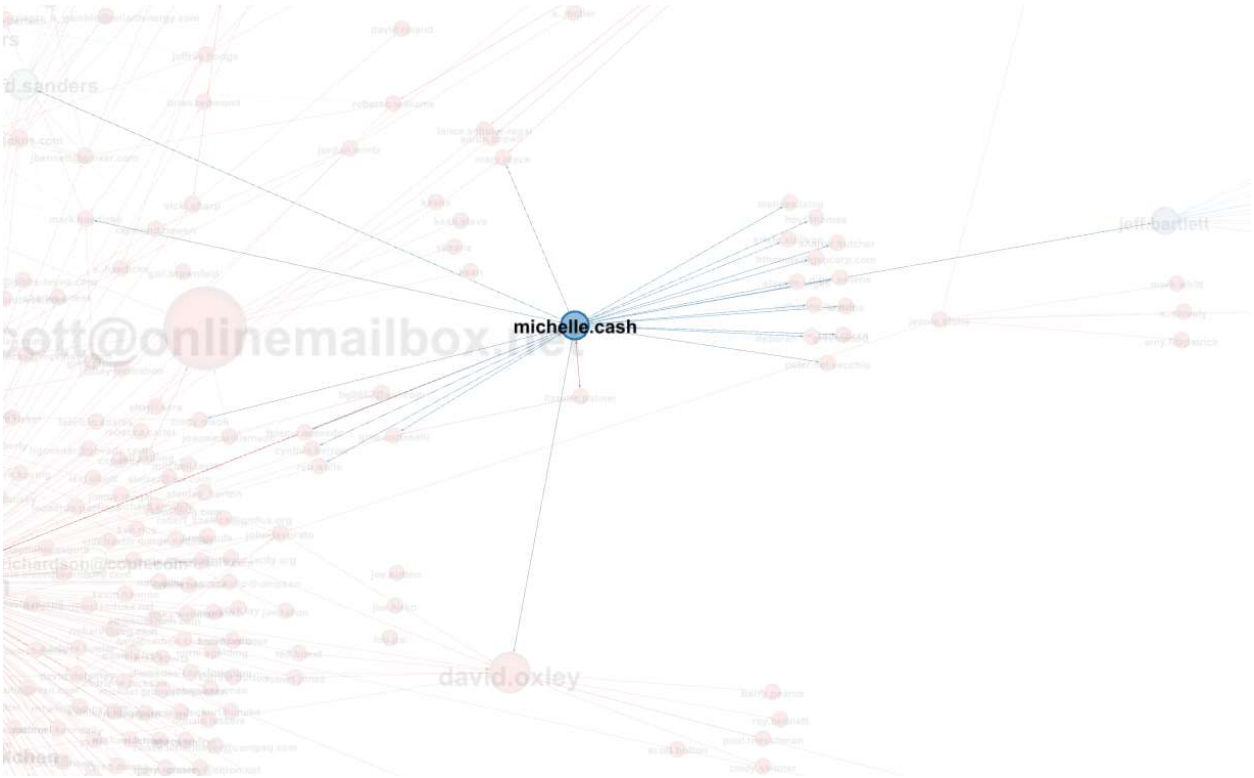
Id	Modularity Class
michelle.cash	4
jeff.bartlett	4
mary.hain	3
tom.hoatson	3
jeff.dasovich	3
rcarroll@bracepatt.com	3
dwatkiss@bracepatt.com	3
gfergus@brobeck.com	3
sarah.novosel	3
richard.sanders	3
mday@gmssr.com	3
tim.belden	3
joe.hartsoe	3
rick.buy	2
rod.hayslett	2
stanley.horton	2
richardson@copn.com	2
phillip.allen	1
john.shelk	1
alan.comnes	1
susan.mara	1
linda.robertson	1
ray.alvarez	1
d..steffes	1
robert.frank	1
b..sanders	1
c..williams	1
kevin.hyatt	0



Class 4 Modularity

In this class there are 2 individuals from the network, Michelle Cash and Jeff Bartlett. Michelle Cash appears to be an employee of Enron while Jeff Bartlett is not. From the email of Michelle Cash it appears she may be a part of Human Resources or employment law. Michelle has a betweenness centrality of 3486.5 which is a high score for this network. It shows as a node that is connected by many short paths throughout the network.

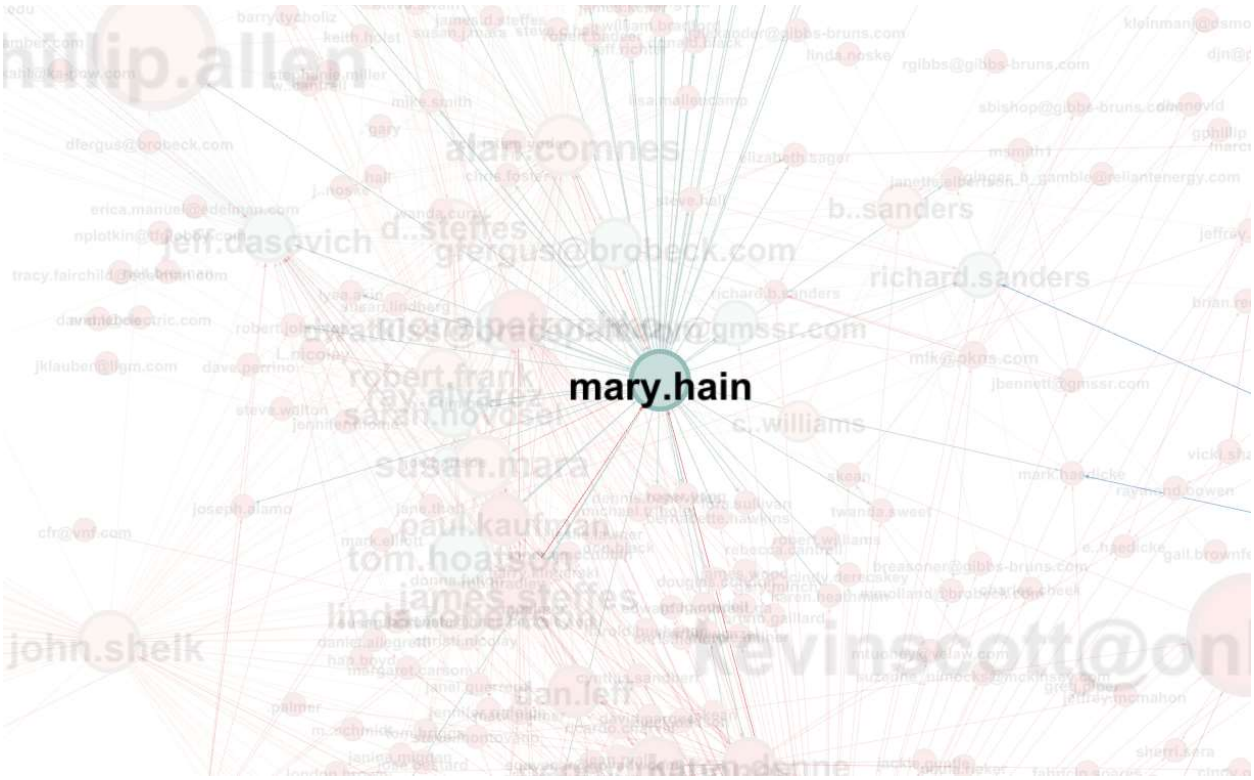
M Cash connects many important clusters but is not central to the main cluster. Possibly a compliance officer or employment law.



Class 3 Modularity

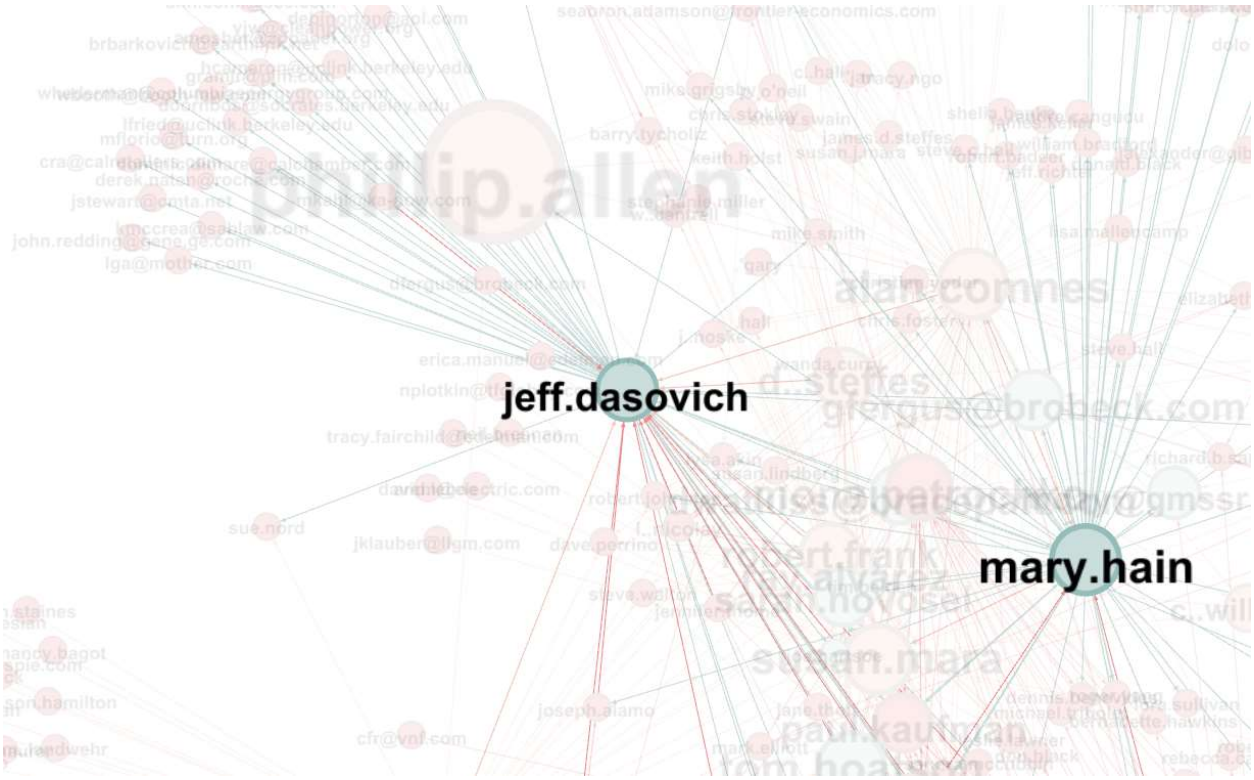
There are 11 individuals in the class. I can retrieve email from 3 employees.

Mary Hain ~ Lobbyist / Connected to a high number of nodes in the network. Probably a part of communicating any changing/updates from the political world.

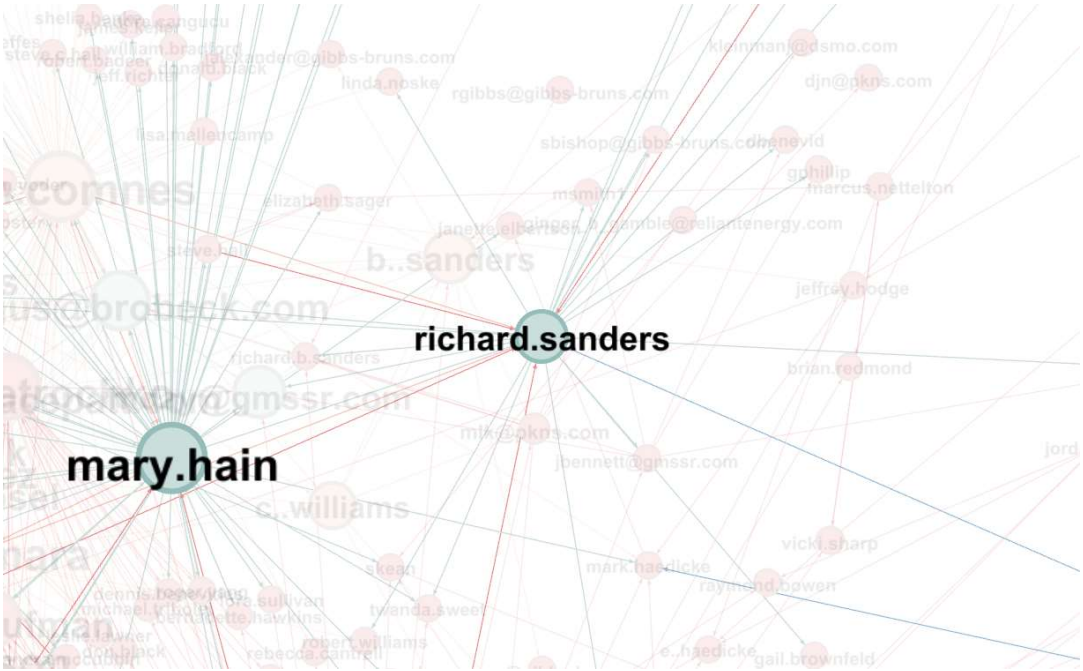


Jeff Dasovich ~ CFO

Important as the head or near head of financial operations. He would not have the same number of contacts as someone in charge of HR but would have many high-level contacts concerning the internal financial position of Enron.



Richard Sanders ~ Law

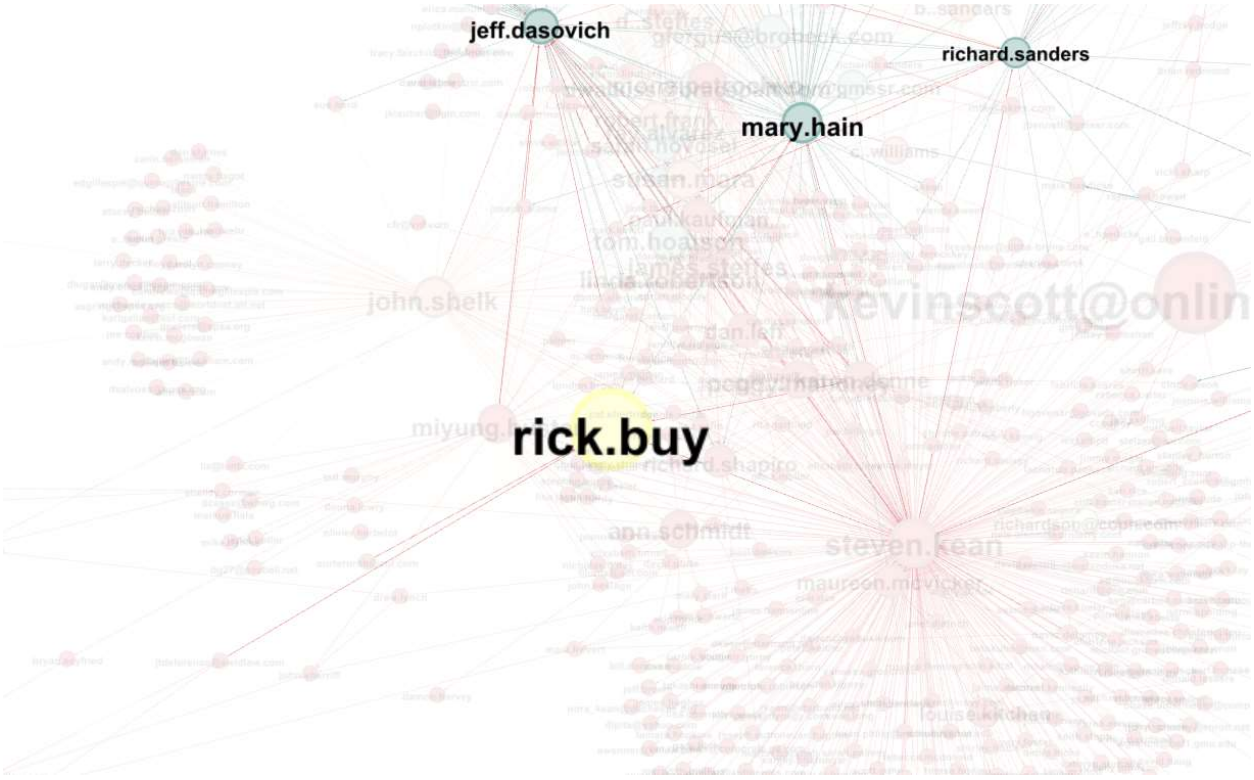


Another corporate attorney at a high level of communication within the network.

Each of these individuals provides an important operating service to Enron. It makes sense that they would be a part of an email chain that discusses topics at the highest level. Each has a high Betweenness Centrality and a moderatley high Closeness Centrality.

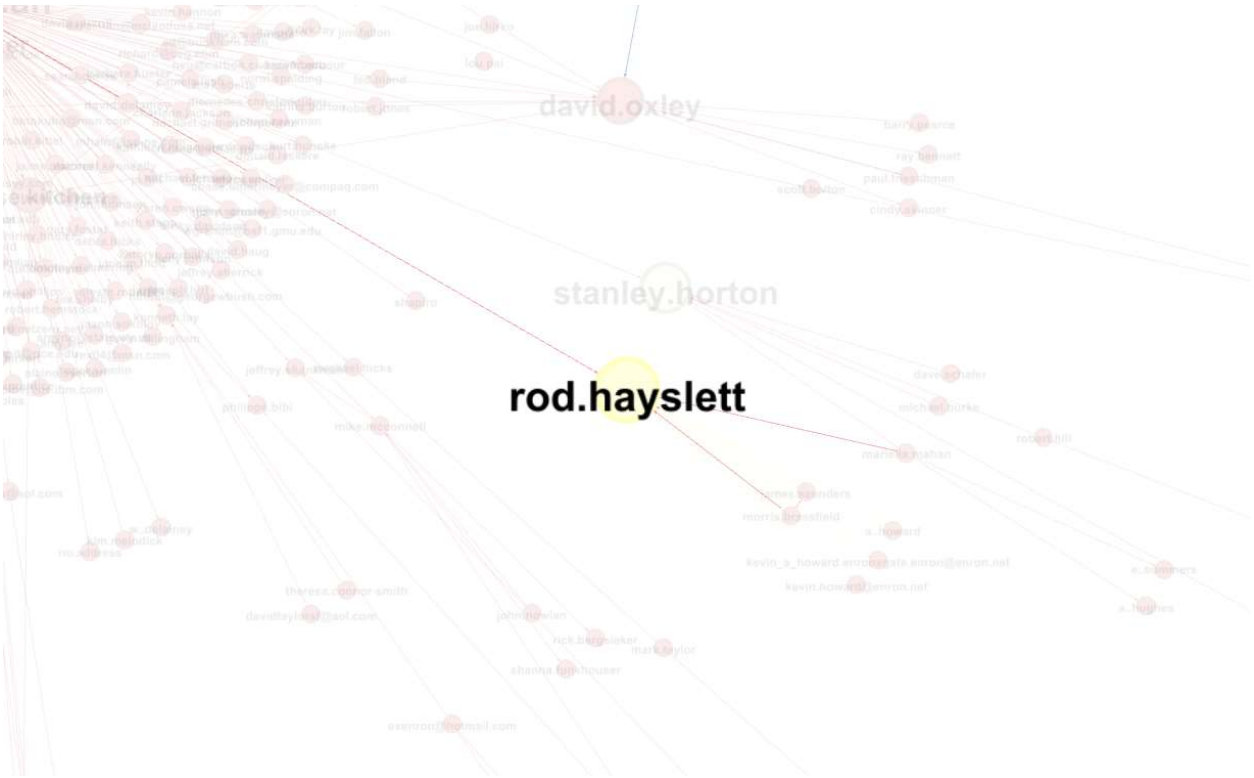
Class 2 Modularity

Rick Buy ~ Risk Management. This individual does not have a high number of direct or indirect contacts but is central to the network. Risk management does not do a lot of social communication as risks are typically held in close confidence with higher management.

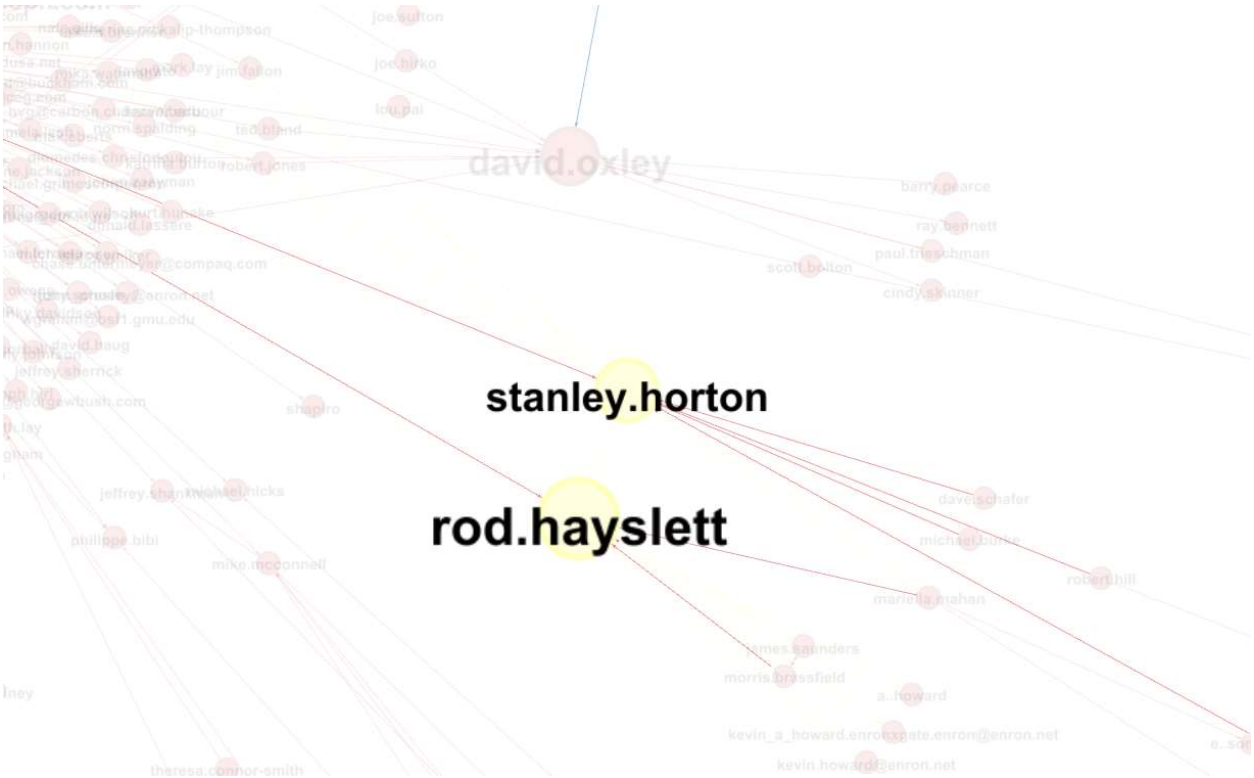


R. Buy has a high closeness Centrality. His emails imply he operated in risk management to some capacity. Yes, this responsibility is highly central to operating a large corporation.

Rick Hayslett ~ High level accountant. High closeness. Similar to risk analysis, accountants don't typically have a high social network. Information is fed up the chain and chief accounts work directly with higher management with information held in close confidence.

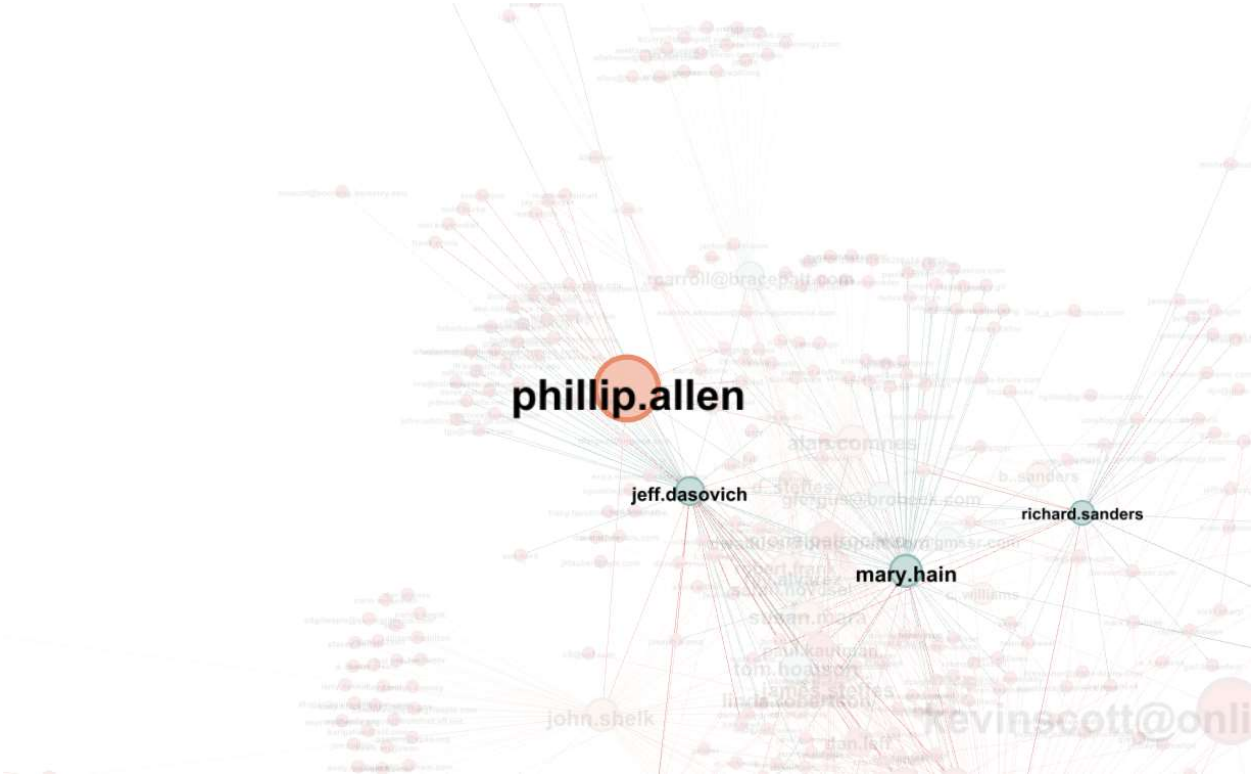


Stanley Horton ~ Corporate lawyer. Medium closeness and low betweenness.

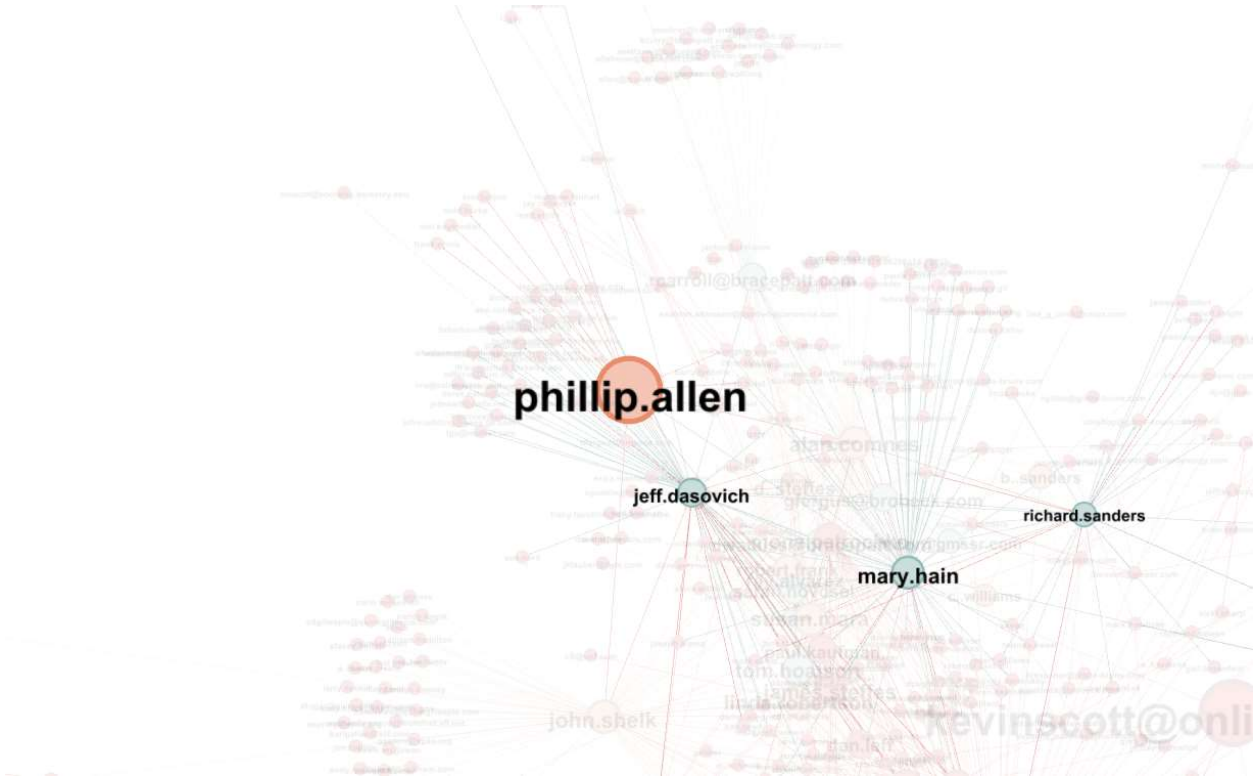


Class 1 Modularity

Phillip Allen ~ Human Resources. High closeness low to mid betweenness. He does not appear to have a high number of connections. Email indicates Human Resources. Could be compliance or some other similar position type.



D Steffes ~ Corporate Lawyer, lobbyist. Low to mid closeness and low to mid betweenness. Appears to be part of the lobbyist arm of the corporation. Centered in the main cluster outside steven.keen.


phillip.allen
jeff.dasovich
richard.sanders
mary.hain