TTS Link Analysis Report
Ruaridh Thomson s0786036

Algorithm implementation

Both PageRank and Hubs and Authorities imagine the graph as a node network where all unique nodes are stored in a dictionary. Nodes correspond to unique emails, where in the case of klay@enron.com and kenneth.lay@enron.com these are treated as separate nodes on the graph. Each node is a Node object that stores the name (email) of the node, the destination links (dest_nodes) of the node (name of other nodes this node points to) and the source links (source_links) of the node (name of nodes pointing to this one). It is possible to quickly (~10 seconds) iterate over graph.txt and populate each node with its destination and source links before performing link analysis. This saves unnecessary iterating to find source or destination nodes during any calculations. All nodes are stored in a dictionary with the node name as the key - this is the graph. The code for the main chunk of each algorithm (pagerank and hubs_auth) follows conventions and implementation outlined in the slides (web-2x2.pdf) and should be readable.

Against the sanity checks, PageRank successfully achieved the correct scores in 10 iterations. Hubs and Authorities (H&S), however, only took 9 iterations to get the values. The number of iterations are defined at the top of the source. Although H&S converged to the sanity checks in 9 iterations, it was observed that calculating hub score before authority score very slightly changed the overall scores compared to authority before hub. With enough iterations this is negligible.

Usefulness

PageRank manages to identify significantly more people who are present in 'roles.txt'. Hubs and authorities identifies what we would expect from both; people with a lot of outgoing and people with a lot of incoming. Though many emails cannot be identified in roles.txt. The companies automatic emailer would be a likely guess as a top hub (effectively spamming, but not spam, everyone with company info).

Is the description 'Employee' just as useful as 'N/A' or 'xxx' or the person not existing in roles.txt. We assume that even if a person is not present in roles.txt that they are an employee. Or are we to assume that because they are not in roles.txt that they may be fake emails or not employed by Enron.

Visualising Key Connections

It is possible to add labels to the connections we are visualising, though as far as I can tell these labels would be the subject of the email - which one would get by reading enron.xml.

The following were chosen purely because they exist in roles.txt and we can observe how information is exchanged by people we know. We are able to observe a variety of employees, all of which rank highly in either PageRank, Hubs or Authorities. We are also able to observe information about the two unknown employees (bill.williams and gerald.nemec). All nodes are included in the graph, showing an interesting divide in communication. See the Appendix for the graph.

Employees graphed:

mark.taylor@enron.com	Employee
gerald.nemec@enron.com	N/A
daren.farmer@enron.com	Manager (Logistics Manager)
sally.beck@enron.com	Employee (Chief Operating Officer)
klay@enron.com	CEO
kenneth.lay@enron.com	CEO (alias of above?)

john.lavorato@enron.com	CEO (Enron America)
kay.mann@enron.com	Employee
kate.symes@enron.com	Employee
pete.davis@enron.com	Used for auto-generated emails (fake user)
bill.williams@enron.com	xxx
geir.solberg@enron.com	Employee (Analyst)
craig.dean@enron.com	Trader
ryan.slinger@enron.com	Trader
albert.meyers@enron.com	Employee (Specialist)

Here we have 15 employees, some of which are obviously important to the company (CEO, Chief Operating Officer, etc.) and some that are of unknown importance. In the case of bill.williams@enron.com, there have been a lot of emails sent from bill and he has been recognised as an authoritative source of information; though it is uncertain who he is.

We will visualise the exchange of information between these individuals. We will use the label to indicate the number of emails.

Some curious observations

pete.davis (the email robot) sends 3615 emails to itself, though this may be to keep record of all unique emails sent.

There is a clear divide with who is sending who emails, with one exception; bill.williams. Though we do not know who he is..

We can see that Kenneth Lay uses klay@enron.com for something (PageRank), but only seems to communicate between these people using kenneth.lay@enron.com.

Implications of the algorithms

This really comes down to what each algorithm achieves. By visualising the results we can easily identify how communication is structured within the company. In the case of bill.williams, he is clearly an important figure - almost central - when observed as not only a hub, but an informed figure. However, we can observe that the only contact he has had from either of the CEOs is once by john.lavorato. While being a well connected person he is not so connected with the owners of the company.

We can see that the left hand side of the graph contains those found by PageRank and the right hand side are found via H&S, suggesting more influential figures in the company are found by PageRank rather than H&S. H&S appears to find a network of people who exchange a lot of information rather than maybe the importance of it. pete.davis may be important here, as the more someone contacts what is the hub email of the company, the more likely they are to be found an authority.

Appendices

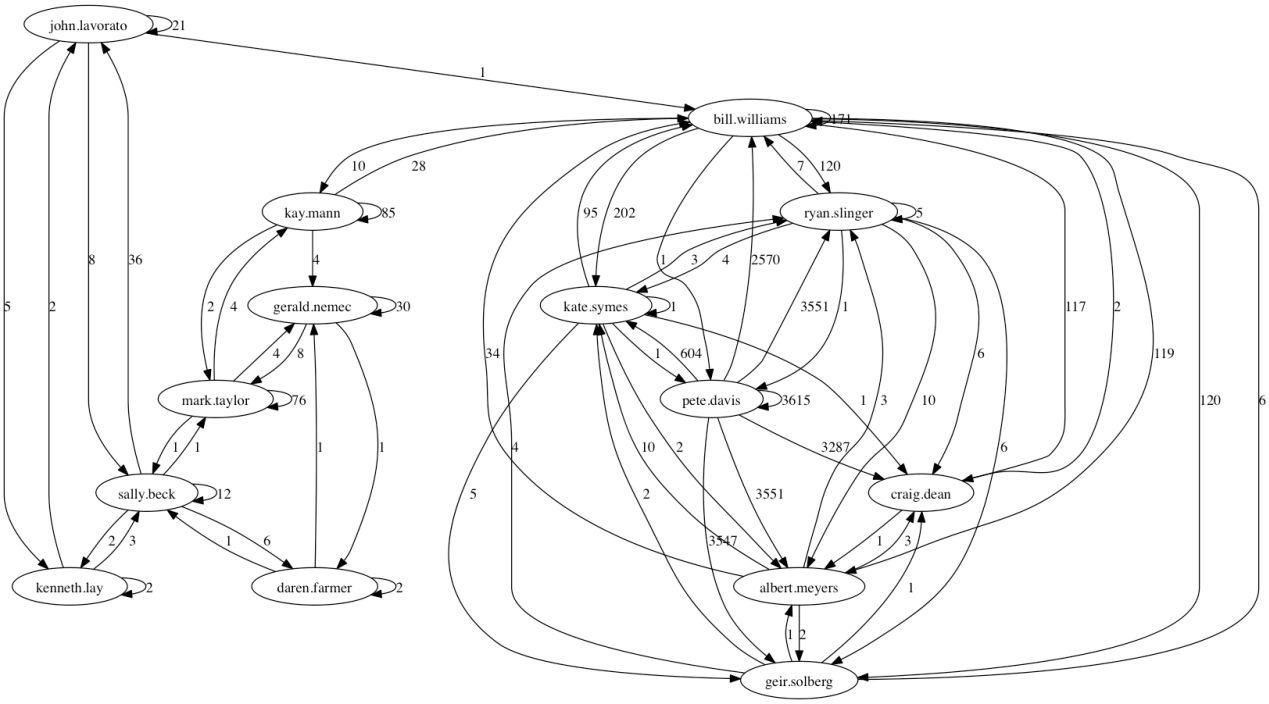
Results

Results from PageRank

0.007376 tana.jones@enron.com N/A
0.00735341 louise.kitchen@enron.com President (Enron Online)
0.00713286 sara.shackleton@enron.com xxx
0.00671554 mark.taylor@enron.com Employee
0.00554463 gerald.nemec@enron.com N/A
0.00530575 daren.farmer@enron.com Manager (Logistics Manager)
0.00459808 sally.beck@enron.com Employee (Chief Operating Officer)
0.00412046 klay@enron.com CEO
0.00405156 john.lavorato@enron.com CEO (Enron America)
0.00390492 kay.mann@enron.com Employee

Results from hubs

0.99928093 pete.davis@enron.com Used for auto-generated emails (fake user)



0.03296957 bill.williams@enron.com xxx
0.01040851 rhonda.denton@enron.com Not in roles.txt
0.00677409 l..denton@enron.com Not in roles.txt
0.00582504 grace.rodriguez@enron.com Not in roles.txt
0.00475687 alan.comnes@enron.com Not in roles.txt
0.00450159 kathryn.sheppard@enron.com Not in roles.txt
0.00401388 kate.symes@enron.com Employee
0.00329035 kysa.alport@enron.com Not in roles.txt
0.00280233 carla.hoffman@enron.com Not in roles.txt

Results from authorities

0.38418728 ryan.slinger@enron.com Trader
0.38417654 albert.meyers@enron.com Employee (Specialist)
0.38384904 mark.guzman@enron.com Not in roles.txt
0.38376442 geir.solberg@enron.com Employee (Analyst)

0.3555813 craig.dean@enron.com Trader
0.277949 bill.williams@enron.com xxx
0.21582593 john.anderson@enron.com Not in roles.txt
0.21574703 michael.mier@enron.com Not in roles.txt
0.1721554 leaf.harasin@enron.com Not in roles.txt
0.14322686 eric.linder@enron.com Not in roles.txt

Graph