# Text Technologies – Exercise 4

s0840449

## 1 Task 1

**PageRank:** Implemented using the iterative-update algorithm. Sink nodes were found as the set-difference of recipients and senders, and corrected for by implicitly connecting them to all nodes in the graph. This gives each node a fraction $\frac{S}{N}$ of the total sink-node PageRank, S, computed once per iteration. The number of out-edges from each node was precomputed for speed.

**HITS:** Implemented using iterative-update. Hub scores were calculated as the sum of the authority scores for in-neighbours, normalised by the magnitude of the hub score vector for all nodes. Authority scores were calculated similarly. Hub scores were updated before authority scores as this gave the smallest error from the sanity values ($2.110 \times 10^{-13}$ as compared to $2.633 \times 10^{-13}$.)

**Observations:** HITS performed poorly on the data-set, likely due to the lack of a query. The top authority scores belonged to traders, who received many emails but rarely replied (Appendix A). The top hub scores belonged to the mailing server (0.999 hub score), and to people who regularly sent many emails such as HR or Administrators (Appendix A.) Neither of the groups are interesting.

PageRank performed better, returning a number of very high ranking employees in the top 10. However the top scoring node, *klay@enron.com*, was a poor choice as it is an alias address with no outgoing emails. As such, it would need to be merged to be interesting.

## 2 Task 2

I began by tackling the problem of duplicated nodes in the graph, i.e. *klay@enron.com* and *kenneth.lay@enron.com*. By using a list of known Enron email aliases[1], I merged duplicate nodes in the graph, so that each node represented one person.

To select 'interesting' nodes I choose to apply PageRank over the reduced graph and take the top 8 results. PageRank was chosen over HITS as it had already been seen to find more interesting nodes. The initial graph was then expanded by adding the second most common correspondent for each node. This was done to find non-direct links between nodes. The second most common correspondent was chosen over the first as the most common correspondents were usually personal secretaries or other non-interesting nodes.

All edges between the selected nodes were then found, and filtered for correspondences which contained at least 5 emails, to reduce graph complexity. Single-neighbour nodes, which were usually uninteresting personal secretaries or colleagues, were then iteratively removed from the graph.

Next, the edges were labelled. I parsed the enron.xml file line-by-line (as it was too badly formed to use a streaming XML parser), and extracted the subject and text of all emails that were represented by an edge in the graph. I initially attempted to adapt *tf-idf* to extract interesting words from the text, but this gave poor results, with many labels being useless words such as 'Best' or 'of'. Instead, I used a simpler approach of taking the most common word from the email subjects, which gave a better but still not totally relevant labelling. In both approaches I filtered words using the NLTK stopwords list[2] and the text of the 'Enron Scandal' Wikipedia page[3], to try and restrict the results to interesting words without manually specifying a list.

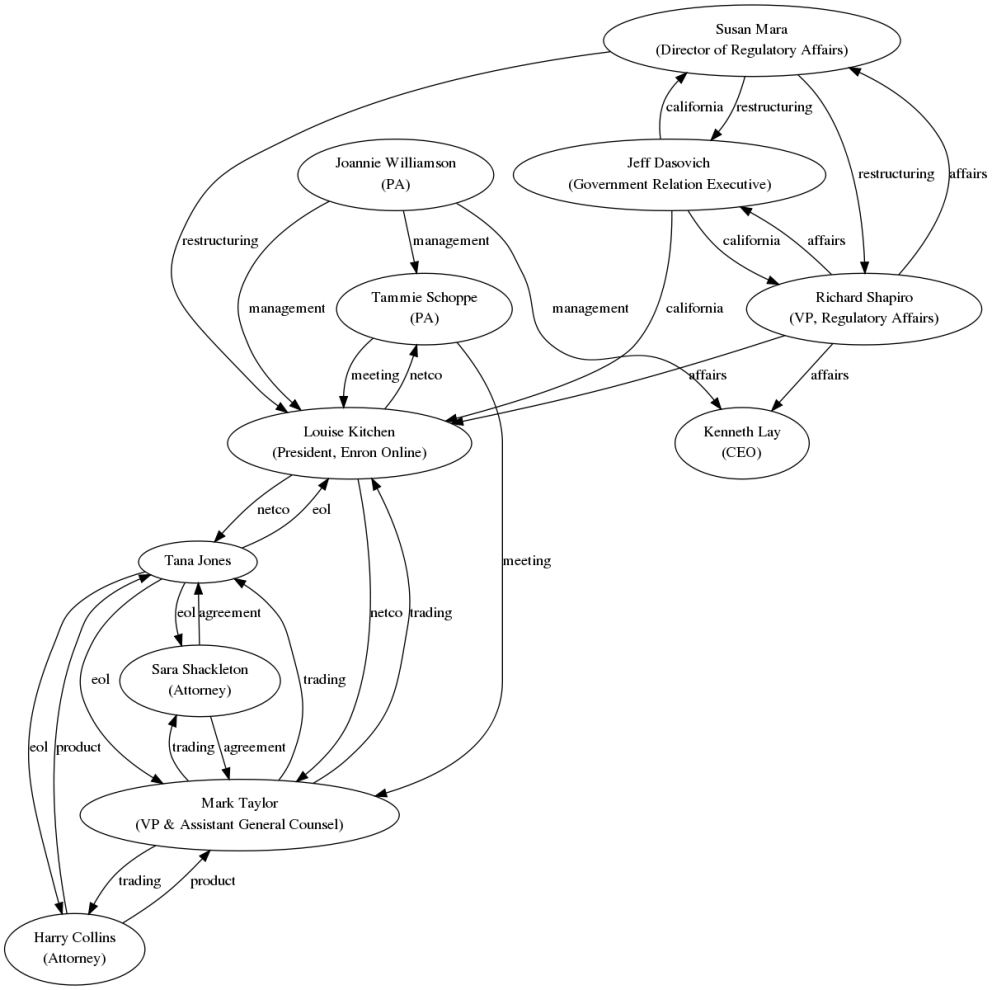


Figure 1: The generated graph.

Finally I attempted to make the graph more human readable. This was done by parsing the roles.txt file to add names and job titles. As this was not a complete list I added details to it for non-matching nodes manually (by searching enron.xml for email signatures/etc.)

The resulting graph (Figure 1) is interesting, but far from perfect. It clearly shows two groups within the selected nodes - high ranking lawyers on one side, and executives on the other. The groups are only connected via the President of Enron Online and some PAs. Unfortunately the graph does a poor job of showing communications between the guilty employees. Only one such employee, Kenneth Lay (CEO), appears on the graph. The conversations around him are interesting as they may indicate regulatory problems and company restructuring, but are not damning. Another wrong-doer, Jefferey Skiller, was in the original graph but was only connected to one node and so was pruned.

The labels are also less than satisfactory. Although the words found are definitely related to Enron ('trading', 'EOL', 'NETCO', etc), they do not directly relate to the scandal, perhaps because the executives did not discuss that area of business often.

# A PageRank/HITS Results Indepth

| Email | Name | Job | Source |
|---|---|---|---|
| klay@enron.com | Kenneth Lay | CEO | roles.txt |
| jeff.skilling@enron.com | Jefferey Skilling | CEO | roles.txt |
| sara.shackleton@enron.com | Sara Shackleton | Attorney | enron.xml |
| tana.jones@enron.com | Tana Jones | ? | N/A |
| mark.taylor@enron.com | Mark Taylor | VP & Assistant General Counsel | [4] |
| kenneth.lay@enron.com | Kenneth Lay | CEO | roles.txt |
| louise.kitchen@enron.com | Louise Kitchen | President (Eron Online) | roles.txt |
| gerald.nemec@enron.com | Gerald Nemec | Attorney | [5] |
| jeff.dasovich@enron.com | Jeff Dasovich | Government Relations Executive | roles.txt |
| sally.beck@enron.com | Sally Beck | COO | roles.txt |

Table 1: PageRank

| Email | Name | Job | Source |
|---|---|---|---|
| ryan.slinger@enron.com | Ryan Slinger | Trader | roles.txt |
| albert.meyers@enron.com | Albert Meyers | Specialist | roles.txt |
| mark.guzman@enron.com | Mark Guzman | Trader | [6] |
| geir.solberg@enron.com | Geir Solberg | Analyst | roles.txt |
| craig.dean@enron.com | Craig Dean | Trader | roles.txt |
| bill.williams@enron.com | Bill Williams | Trader | [7] |
| john.anderson@enron.com | John Anderson | Trader | enron.xml |
| michael.mier@enron.com | Michael Mier | ? | N/A |
| leaf.harasin@enron.com | Leaf Harasin | Trader | [8] |
| eric.linder@enron.com | Eric Linder | Specialist | [9] |

Table 2: Authority Scored

| Email | Name | Job | Source |
|---|---|---|---|
| pete.davis@enron.com | N/A | Broadcast Proxy | roles.txt |
| bill.williams@enron.com | Bill Williams | Trader | enron.xml |
| rhonda.denton@enron.com | Rhonda Denton | Lawyer | enron.xml |
| l..denton@enron.com | Rhonda Denton | Lawyer | enron.xml |
| grace.rodriguez@enron.com | Grace Rodriguez | HR | enron.xml |
| alan.comnes@enron.com | Alan Comnes | Director | [10] |
| kathryn.sheppard@enron.com | Kathryn Sheppard | Administrator | enron.xml |
| kate.symes@enron.com | Kate Symes | Clerk | enron.xml |
| kysa.alport@enron.com | Kysa Alport | Deal Control | enron.xml |
| carla.hoffman@enron.com | Carla Hoffman | Trader | enron.xml |

Table 3: Hub Scored

## References


[1] Social Network Analysis Tool (SNAT).
http://snat.googlecode.com/svn-history/r740/trunk/snat/snat/Resources/emails

[2] NLTK Stopwords Corpus (Porter et al).
http://nltk.googlecode.com/svn/trunk/doc/book/ch02.html

[3] Wikipedia Enron Scandal page.
http://en.wikipedia.org/wiki/Enron_scandal

[4] Mark Taylor's LinkedIn page.
http://www.linkedin.com/pub/mark-taylor/11/8b6/507

[5] Witness Statement from Gerald Nemec (Preliminary Presentation.)
http://datasets.opentestset.com/datasets/Enron_files/full/nemec-g/

[6] Enron Employee Status Report.
http://www.isi.edu/ adibi/Enron/Enron_Employee_Status.xls

[7] NY Times Article referring to Bill Williams.
http://www.nytimes.com/2005/02/04/business/worldbusiness/04iht-enron.html?_r=0

[8] Leaf Harasin's LinkedIn page.
http://www.linkedin.com/pub/leaf-harasin/6/331/3b5

[9] Eric Linder's LinkedIn page.
http://www.linkedin.com/pub/eric-linder/7/578/837

[10] Alan Colmnes' LinkedIn page.
http://www.linkedin.com/pub/alan-comnes/5/242/300