Text Technologies – Assignment 4 – Report **matric:** s0837795

I. PageRank and HITS

The PageRank and HITS are standard implementations of the algorithms described in the lecture slides. Both store the graph.txt in memory as two dictionaries – one each for incoming and outgoing edges of the graph from the point of each vertex – e.g. the relations A->B, A->C, A->B would be represented by the outgoing graph as the entry with key A, values in the list ['B','C','B'] – {'A': ['B','C','B']}, whilst for the incoming graph these would be {'B':['A','A']}, {'C':['A']}. This is different for the HITS implementation, where I use a dictionary of dictionaries, instead of a dictionary of lists, and I also explicitly store the sums explicitly in the dictionary, which helps improve performance. The rest of the implementation just follows the algorithms on the slides and Wikipedia.

The results from both algorithms allow us to make some interesting observations. For PageRank:

1. 0.00232272 klay@enron.com
6. 0.00079018 kenneth.lay@enron.com

Both (1) and (6) are high PR scores that significantly influence the PR scores of linking nodes in the graph, but as one can easily observe both emails represent the same person. If one further explores the dataset, one will quickly realize there are many more instances where multiple nodes represent the same person, due to the fact that the data has not been pre-processed to remove such abnormalities. In addition, both nodes represent the CEO of Enron and the person most involved with the fraud.

For the HITS algorithm, one can notice that the node pete.davis@enron.com has a very large hubs score - 0.99928093, substantially larger than the node with the second largest score in the graph - 0.03296957 for bill.williams@enrron.com. This score implies that pete.davis@enron.com has sent emails to almost every person in the graph, which is consistent with the reported role of that account as a broadcast proxy for auto-generated emails inside the company.

II. Influential people, PageRank, HITS and visualization

For this task, I decided to first clean up the dataset. As discussed earlier, there is a duplicate entry in the list of highest 10 PR scores and many more duplicates in the rest of the ranking. These affect relationships and could cause one to miss an important connection between two people or to artificially inflate the PR score of a node. For example, if both klay@enron.com and kenneth.lay@enron.com forward to the same mailbox, but klay@enron.com is the default identity, kenneth.lay@enron.com would have a large PR score because it receives a lot of email (a lot of nodes point to it), but sends out little (if any) emails (it points to few nodes), yet it all relates to the same person. The opposite is also possible – a person who uses multiple mailboxes for different purposes. In this case I would like to know what these purposes are and what is the flow of information between Enron's top executives, which is why for this task I will consider entries such as klay@enron.com, Kenneth.lay@enron.com, kenlay@enron.net, etc. as one node on the graph.

To judge the effectiveness of PageRank, and Hubs and Authorities scores in determining influential people inside Enron, I decided to manually compile a list of the company's top 13 executives - Kenneth Lay, Jeffery Skilling, Vince Kaminski ,Louise Kitchen ,Jeff Dasovich ,Steven Kean, Sally Beck,John Lavorato, Richard Shapiro, Jeffrey Shankman, Rod Hayslett, Rick Buy, David Delainey. Next, I ran the algorithms on the improved dataset and obtained the scores summarized in Figure 1.

PageRank appears to be very good at identifying influential people inside the company – in the list of the highest 15 PR scores, 9 are in the list of 13 top ranking Enron executives. In contrast, for HITS, the top 15 authority and hubs scores contain none of the 13 top executives. Nevertheless, the first five authority scores are nodes which represent Traders. Traders, due to the nature of their role, often receive substantial amount of email communication – research reports, market reports, P&L updates, price quote requests, orders, etc., but reply to few of them via email – instead they would call up clients on the phone. This would explain their large authority scores.

Going back to PageRank, in the top 10 list of people by PR scores, one can observe that 4 of them have no title and do not appear to be in executive positions, yet have very large PR scores. I have amended these to Figure 1 in yellow. These are the so-called "concealed relationships" described by Nishith Patha and Jaideep Srivastava [1] and I will be interested to explore the connection between these actors and the top executives in the company. To do that, I build a graph using the following algorithm on the pre-processed dataset:

- Each link is from node A to node B, such that A and B are both in the table in Figure 1
- A directed link from node A to node B (A -> B) implies more than 20 emails sent from A -> B
- Each link A -> B is unique, e.g. 40+ emails do not count as two links

The visualisation shows a strong pairwise relationship between Mark Taylor and Sara Shackleton, Mark Taylor and Louise Kitchen (President Enron Online), Sara Shackleton and Tana Jones, Tana Jones and Louise Kitchen (President Enron Online), Tana Jones and Sara Shackleton. It appears that Mark Taylor, Sara Shackleton, Mark Taylor and Louise Kitchen work together. Although no data is available to verify this proposition, it appears that these individuals are all part of Enron Online and are likely a closely knit team.

Interestingly, due to the restriction of 20 emails or more from A -> B, Gerald Nemec does not appear in the visualization. This would imply that his high PR score derives from sparser connections to the wider set of individuals in the table below. This is similar for Kenneth Lay, who only has a single incoming connection from Steven Kean. This, however, is quite natural for a CEO of a large organisation – he would receive a lot of emails, but reply to few of them.

III. Summary

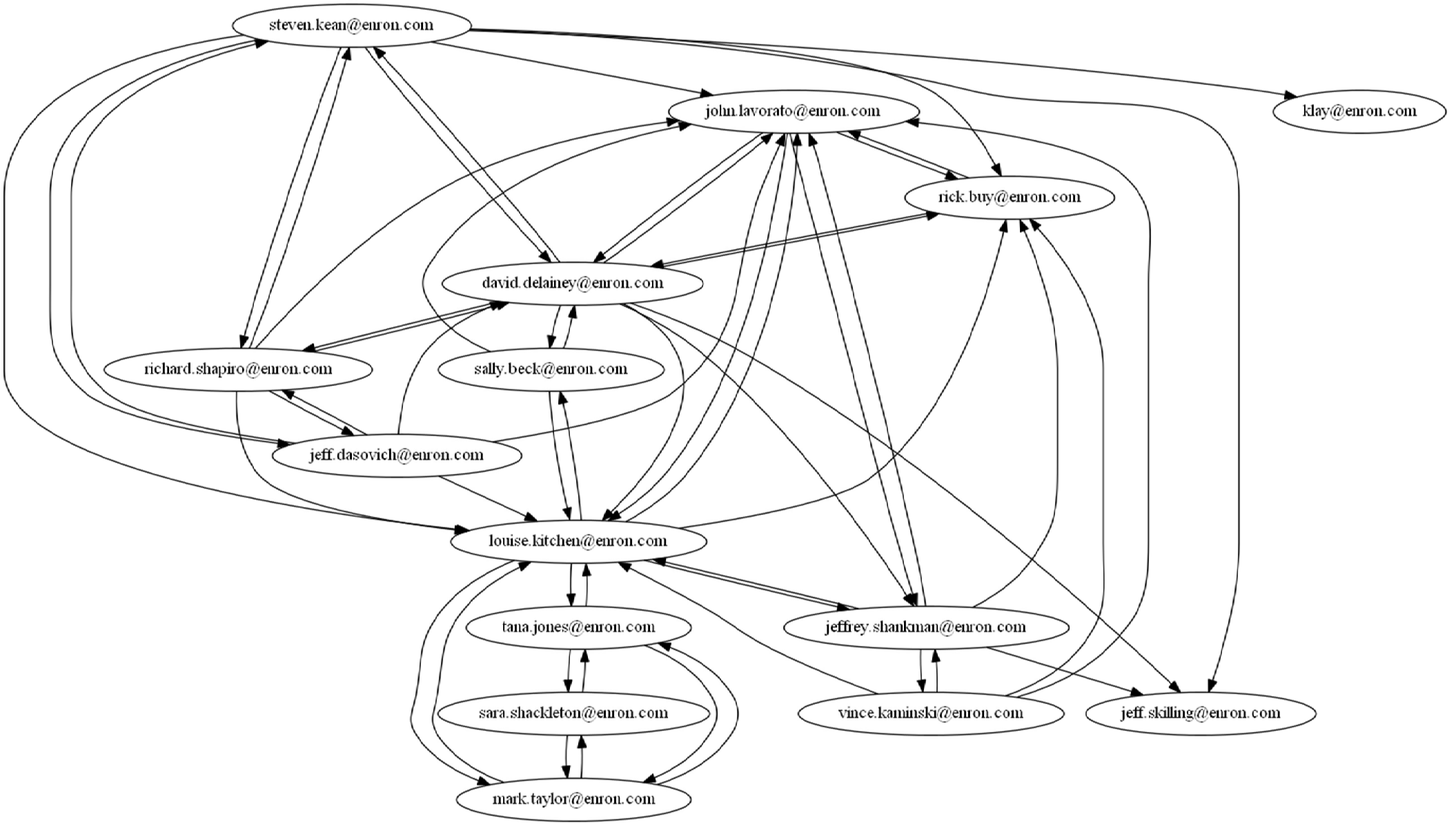
- Ran HITS, PR algorithm on cleaned up data and wrote a graph construction algorithm to find key connections
- PageRank is much more effective than HITS for finding influential people on this particular dataset
- However, it is not a 'magic bullet' – David Delainey, CEO of Enron North America ranks 105th.
- Also exposes interesting relationships between non-executives, could be improved with clustering.
- Helped us identify key staff of Louise Kitchen (President of Enron Online)

Name	Role	PageRank	PR #	Auth	Auth #	Hub	Hub #
Kenneth Lay	CEO and chief fraudster	0.00317386	1	3.51E-06	1024	2.04E-05	425
Jeffery Skilling	CEO	0.00093598	2	3.51E-06	1021	3.50E-07	1128
Sara Shackleton	N/A	0.00091693	3	1.49E-05	492	1.49E-05	477
Vince Kaminski	Risk Management Head	0.00091166	4	5.10E-06	856	1.54E-06	759
Tana Jones	N/A	0.0008914	5	4.44E-05	278	4.06E-05	304
Mark Taylor	Employee	0.00085095	6	1.88E-05	473	6.34E-06	570
Louise Kitchen	President Enron Online	0.00083256	7	6.37E-05	228	2.02E-05	428
Jeff Dasovich	Government Relation Exec	0.00072749	8	0.0002107	43	0.00101	32
Gerald Nemec	N/A	0.00071858	9	5.70E-06	798	2.10E-06	717
Steven Kean	Chief of Staff	0.00064583	10	0.0001327	62	2.10E-05	422
Sally Beck	Chief Operating Officer	0.00057869	11	1.24E-05	528	4.94E-06	598
John Lavorato	CEO enron america	0.00057077	12	4.09E-05	295	0.00013	148
Richard Shapiro	Regulatory Affairs	0.00053977	13	0.0001881	44	1.67E-05	434
Jeffrey Shankman	Enron Global Mkts	0.00045763	16	5.87E-06	781	9.30E-07	843
Rod Hayslett	CFO and Treasurer	0.0003458	26	4.80E-07	3569	3.10E-07	1198
Rick Buy	Chief Risk Mngment Officer	0.0003234	28	3.91E-06	934	1.88E-06	736
David Delainey	Ceo Enron North America	1.17E-04	105	1.53E-05	489	2.59E-05	404

Figure 1 - The set of top 13 executives and top 10 highest PR scores (from pre-processed data, not original data). In yellow, non-executives with curiously high PR scores.

References:

[1] Nishith Pathak and Jaideep Srivastava, Automatic Extraction of Concealed Relations from Email Logs, URL: http://vw.indiana.edu/netsci06/conference/Pathak_Automatic.pdf


steven.kean@enron.com
john.lavorato@enron.com
klay@enron.com
rick.buy@enron.com
david.delainey@enron.com
richard.shapiro@enron.com
sally.beck@enron.com
jeff.dasovich@enron.com
louise.kitchen@enron.com
tana.jones@enron.com
jeffrey.shankman@enron.com
sara.shackleton@enron.com
vince.kaminski@enron.com
jeff.skilling@enron.com
mark.taylor@enron.com