Text Technologies for Data Science Assignment 4

s1140740

November 17, 2014

PageRank Algorithm

People are represented as nodes with a directed edge from a person sending an email to a person receiving the email. The weight (w) of this edge represents the number of emails the person sent. PageRanks (PR) are initialised to $1/N$ where N is the total number of people. 10 iterations of the algorithm are then executed. Each iteration S, the sum of PageRanks of sink nodes, is computed and people's PageRanks are updated using the PageRank formula.[1]

[1]
$$PR(x) = \frac{1-\lambda+\lambda S}{N} + \lambda \sum_{y \to x} \frac{w \cdot PR(y)}{out(y)}$$

After running the algorithm for 10 iterations, `klay@enron.com` turned out to be the email address with the highest PageRank of 0.008027. Moreover, the email address in 6th place is `kenneth.lay@enron.com` which is probably the email address of the same person, CEO of Enron, Kenneth Lay. We can thus conclude that the PageRank algorithm did quite well in figuring out who the influential people are. The top 5 PageRanks can be found in table 1.

Email	PageRank
klay@enron.com	0.008027
jeff.skilling@enron.com	0.003019
sara.shackleton@enron.com	0.002961
tana.jones@enron.com	0.002855
mark.taylor@enron.com	0.002753

Table 1: Top 5 PageRanks

HITS Algorithm

Same representation as for PageRank is used. However, hub and authority values are initialised to $1/\sqrt{N}$. Each iteration the hub value of a node is updated using authority values of nodes it is pointing to[2] and the authority value is updated using hub values pointing to it.[3] Both hub and authority values are then normalised.[4]

[2]
$$H(x) = \sum_{y \leftarrow x} A(y)$$

[3]
$$A(x) = \sum_{y \to x} H(y)$$

[4]
$$\sum_x H(x)^2 = 1 = \sum_x A(x)^2$$

After running the algorithm for 20 iterations (required to reach values in the sanity check), `pete.davis@enron.com` turned out to be

the email address with the highest hub score of 0.999281. This is consistent with his role description in `roles.txt` - broadcast proxy for auto-generated emails. Other than that there were no abnormalities. Email addresses with top 5 hub scores and top 5 authority scores are in tables 2 and 3, respectively.

Email	Hub score
pete.davis	0.999281
bill.williams	0.032970
rhonda.denton	0.010408
l..denton	0.006774
grace.rodriguez	0.005825

Table 2: Top 5 Hub scores

Email	Auth score
ryan.slinger	0.384187
albert.meyers	0.384177
mark.guzman	0.383849
geir.solberg	0.383764
craig.dean	0.355581

Table 3: Top 5 Auth scores

Visualizing key connections

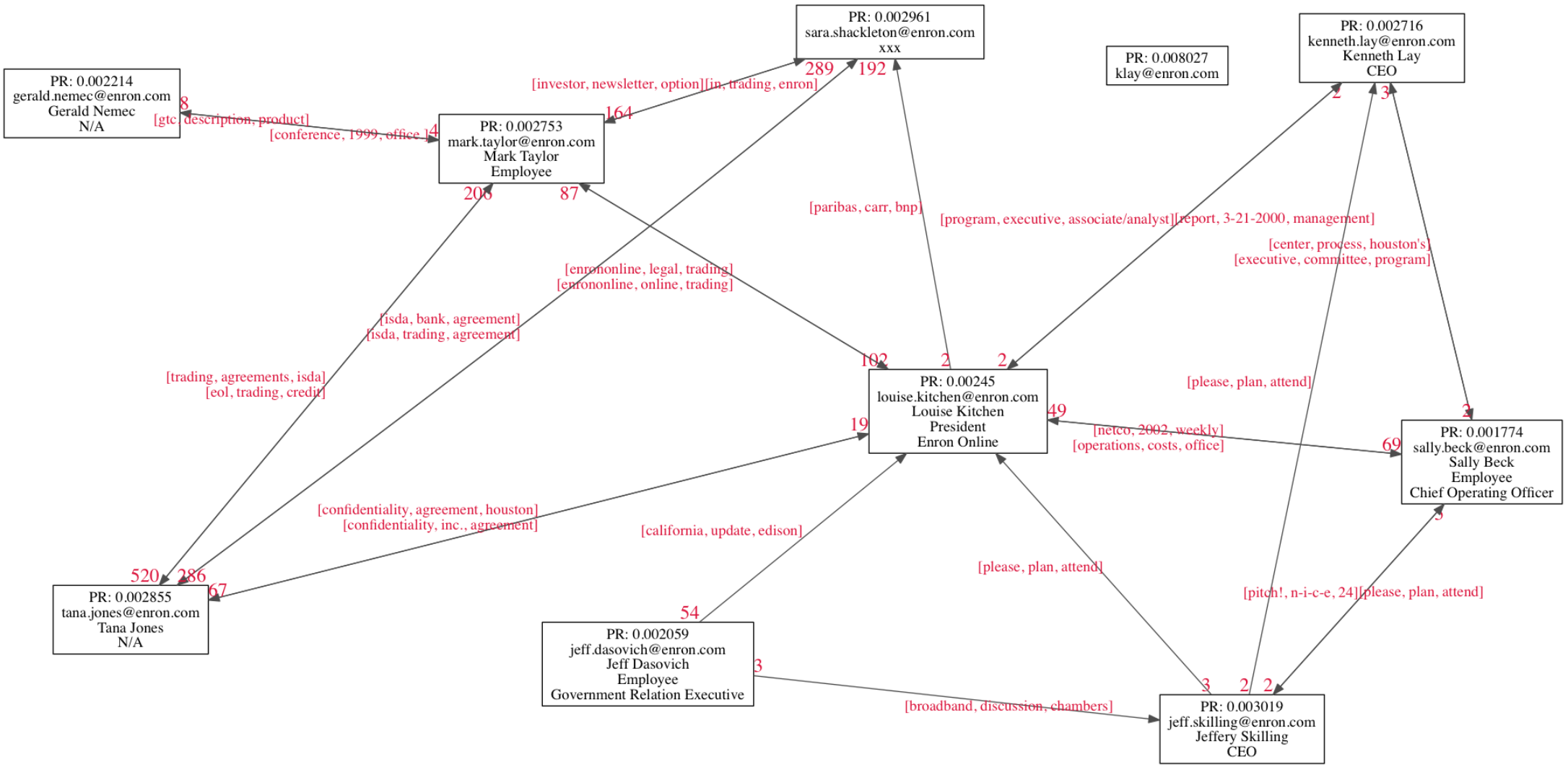
I visualised connections between 10 people with the highest PageRank. It seemed that the algorithm ranked people with great influence highly and thus visualising their connections could provide some additional information about the scandal. I used the networkx library[5] to create a graph of these connections which is then outputted into a file `graph.dot` and visualised with graphviz.[6]

I used information from `roles.txt` to assign names and roles to email accounts of Enron employees. These are displayed, together with PageRank in the box of a node. Furthermore, if the number of emails exchanged between two people is greater than 1 it is displayed at the tail of an edge in red colour and 3 words with the highest frequency in emails between two nodes are displayed as a label for an edge. Stop tokens such as `re:`, `fw:`, `&`, `to`, `of`, `-`, `and`, `for` and `the` are not considered as they are common to many emails.

My algorithm does not require any additional manual tuning which means that it is very automatic.

[5] NetworkX, https://networkx.github.io

[6] graphviz, http://www.graphviz.org


PR: 0.002214
gerald.nemec@enron.com
Gerald Nemec
N/A
PR: 0.002961
sara.shackleton@enron.com
xxx
PR: 0.008027
klay@enron.com
PR: 0.002716
kenneth.lay@enron.com
Kenneth Lay
CEO
PR: 0.002753
mark.taylor@enron.com
Mark Taylor
Employee
PR: 0.00245
louise.kitchen@enron.com
Louise Kitchen
President
Enron Online
PR: 0.001774
sally.beck@enron.com
Sally Beck
Employee
Chief Operating Officer
PR: 0.002855
tana.jones@enron.com
Tana Jones
N/A
PR: 0.002059
jeff.dasovich@enron.com
Jeff Dasovich
Employee
Government Relation Executive
PR: 0.003019
jeff.skilling@enron.com
Jeffery Skilling
CEO
8
[gtc, description, product]
[conference, 1999, office]
4
289
192
[investor, newsletter, option]
[in, trading, enron]
164
208
87
[paribas, carr, bnp]
[program, executive, associate/analyst]
[report, 3-21-2000, management]
2
3
[center, process, houston's]
[executive, committee, program]
[enrononline, legal, trading]
[enrononline, online, trading]
[isda, bank, agreement]
[isda, trading, agreement]
[trading, agreements, isda]
[eol, trading, credit]
102
2
2
[please, plan, attend]
49
19
[netco, 2002, weekly]
[operations, costs, office]
69
3
[confidentiality, agreement, houston]
[confidentiality, inc., agreement]
[california, update, edison]
[please, plan, attend]
[pitch!, n-i-c-e, 24]
[please, plan, attend]
520
286
67
54
3
3
2
2
[broadband, discussion, chambers]