CS5234J – Final Project Report

In the final project report, you will use the code you have developed for the final project to answer questions about your implementation and analyse basic properties of the network induced by the full Enron Email dataset.

What and how to submit Create a PDF file with the answers for the questions below, and submit it on Moodle.

Submission deadline: 10:00am, 9 August 2021

Question 1 (60%)

Pick one of the functions among those you implemented for the final project, and answer the following questions. If you worked on the project jointly with another student, you and your partner must choose **different** functions to discuss.

1. Include the Python code for your chosen function, and **briefly** describe its implementation highlighting any challenges and pitfalls you had to deal with. (**20%**)

2. Assume the `count()` action is called on the RDD returned by the function you have chosen. Draw a *lineage graph (DAG)* capturing the series of transformations triggered by the call. Your diagram should include dependencies between individual RDD partitions. You can assume that all RDDs have exactly *two* partitions. (**30%**)

3. Are there any *narrow* dependencies in the lineage graph you have drawn in Question 1.2? Give one example of a narrow dependency if it is present. (**5%**)

4. Are there any *wide* dependencies in the lineage graph you have drawn in Question 1.2? Give one example of a wide dependency if it is present. (**5%**)

5. Using the lineage graph from Question 1.2, identify the *stage(s)* and *task(s)* of your pipeline. (**10%**)

Question 2 (30%)

Many complex networks induced by either human interaction (e.g., friendship or follower graphs in online social networks, or World-Wide Web) or natural phenomena (e.g., protein-to-protein interaction networks) are known to be *scale-free*, i.e., the degree distribution in such networks follows a *power law*

$$p(k) \sim k^{-\alpha}, \alpha > 1, \tag{1}$$

where k is a non-negative integer, and $p(k)$ is the probability that a node has the degree k.

Observe that (1) implies that a power law distribution will look roughly as a straight line if graphed on a log-log scale with the line's slope being equal to its exponent α.

A remarkable feature of a power law distribution is that all its moments above 1 are infinite, which implies that fluctuations around the mean can be arbitrarily high. This means that when we randomly choose a node in a scale-free network, we do not know what to expect: The selected node's degree could be tiny or arbitrarily large. In particular, this property predicts the emergence of *hubs*, i.e., nodes whose degrees can become disproportionally large (think e.g., of the number of people following a celebrity account on Twitter).

Pick a consecutive 12 months period contained within the range from **January 2000** to **March 2002** (inclusively). Use the functions you implemented for Tasks 1 and 3 of the final project to extract a *slice* of the weighted network contained within the period you chose. Analyse its properties by answering the questions below. Note that if you worked on the final project jointly with another student, you and your partner must choose **different** slices to analyse.

1. The number of connections originating at or attracted by nodes in a scale-free network follows an *80/20 rule*, that is, roughly 20% of the nodes are either origins or destinations of 80% of all edges in the network. (These 20% of nodes are, in fact, the hubs.)[1] Use the code you developed for Task 4 of the final project to compute the total weighted degree of all edges originating at (respectively attracted by) the 20% highest out-degree (respectively, in-degree) nodes in the network slice you selected. Briefly describe the methodology you used and your findings. Does the 80/20 rule indeed apply to your chosen network slice? You may use visualisations to support your conclusions. (**15%**)

2. Another interesting property of the scale-free networks (colloquially known as "the rich get richer") is that the maximum node degree k_{max} is *directly proportional* to the number of nodes in the network. Use the functions you implemented for Tasks 3 and 4 of the final project to compute k_{max} (for both in and out degrees) and the number of nodes for the sub-slices containing the Emails sent within the first n months, where $1 \leq n \leq 12$, of your chosen network slice. Briefly describe the methodology you used and your findings. Does k_{max} for either in or out degrees (or both) indeed grow linearly with the number of nodes in your chosen network slice? You may use visualisations to support your conclusions. (**15%**)

[1] This rule was originally formulated by a 19th century economist Vilfredo Pareto in the context of wealth distribution in a capitalist free-market society: Roughly 80% of money is earned by only 20% of the population.