PFP Project Proposal
Jeremy Carin

Background:
The PageRank algorithm is a widely used method for ranking the importance of web pages in search engines. It was developed by Google co-founders Larry Page and Sergey Brin as a way to measure the importance of web pages by analyzing the link structure of the web. The PageRank uses the link structure of the web to measure the importance of web pages. The idea behind this approach is that a web page with many high-quality incoming links from other reputable web pages is likely to be more relevant and useful than a web page with fewer or lower-quality incoming links.

Implementation:
The core implementation is a Map-Reduce problem, which makes it highly parallelizable. The input data consists of the web graph, represented as a set of nodes (web pages) and edges (links between pages). Each map task is responsible for processing a portion of the input data, which may consist of one or more nodes and their associated edges.

The map function for the PageRank algorithm computes the PageRank score for each node in its portion of the input data, and produces a set of intermediate key-value pairs as output. The key in each intermediate key-value pair is the node ID, and the value is the PageRank score for that node.

The reduce function for the PageRank algorithm receives a group of intermediate key-value pairs with the same node ID and combines them into a single output value, which is the final PageRank score for that node.

The MapReduce framework then repeats the map and reduce stages for a specified number of iterations, at which point the final PageRank scores for all nodes in the input data are produced as output.

Another feature of PageRank is the dampening factor, which is used to prevent "rank sinkholes," which are web pages that have a high PageRank score but do not pass that score on to other pages through their outgoing links. Without the damping factor, these sinkholes could potentially dominate the PageRank scores of other pages, leading to an uneven distribution of scores.

The damping factor works by reducing the influence of incoming links on a page's PageRank score. It does this by "dampening" the contribution of each incoming link, so that a page's PageRank score is not solely determined by the number and quality of its incoming links.

I plan to first develop a general parallel Map-Reduce framework, and then create a method to load a large amount of data and process it parallelly. To create the sequential version that I compare it against, I will simply remove the parallelism from the Map-Reduce framework – a single-threaded Map-Reduce. I may also try to implement a sequential version of the PageRank algorithm without any map-reduce.