## SUMMARY & PROBLEM DESIGNED

The web is a network of webpages and links between them where the nodes are the webpages and the edges are the hyperlinks. Here we perform a data and network analysis of the Cloud Computing and Distributed Computing hyperlinked Wikipedia Pages to gain insights about the relationships between the contents and topics (hyperlinks). These hyperlinks are going to be unique to Cloud Computing or Distributed Computing or common to both Wikipedia pages.

## RESEARCH DESIGN, MEASUREMENT & NETWORK METHODS

To facilitate the data and network analysis, the following graphs were generated for analysis: Directed graph, Co-Linked Undirected graph, Co-Linking Undirected graph and the largest strongest Connected Component of Directed graph. Some of the node labels overlapped hence 2 plots were created with 2 less overlapping nodes to make easier to identify and read.
In additional. Also the way we got the plots to look reasonable to read the labels and the corresponding nodes was to tweak the edge width and also tweak the y and x offsets. You can adjust number from 0.003. the “edge_width1 = [0.003*e for e in edge_width1]” was used to account for the node label length in the x offset. In addition, node_color was used to differentiate between the "top_nodes1" & top_nodes2 from remaining nodes "rem_nodes1" & "rem_nodes2"

## OVERVIEW OF PROGRAMMING WORK

For any diagnostic, descriptive and statistical analysis to be performed, there must be available data. Critical for this analysis of the Cloud Computing and Distributed Computing hyperlinked Wikipedia Pages, the following python modules the was performed using the Wikipedia, Networkx, Pygraphviz and community modules were used. Wikipedia for Collection of the Cloud Computing and Distributed Computing hyperlinks. Networkx for analysis and computation of

graphs. Pygraphviz for Access to Graphviz graph data structure to create graphs. Community partitioning of the graph nodes in to communities for analysis.

## RESULTS

Networkx computation of diagnostics and descriptive statistics provided these insights:

**Directed Graph:** Count of nodes & edges were: Cloud Computing (396 hyperlinks & 21191 edges) and Distributed computing (341 hyperlinks & 17250 edges), 118 hyperlinks common to both Wikipedia pages and a combined total of 619 nodes and 29955 edges. Connectivity type was a Simple, unweighted, graph, weakly connected directed graph with a reciprocity of 0.5244. It's not bipartite graph, not a tree, has no isolates and not strongly connected and has 61 strongly connected components. Centrality Indices computed were (out_degree, in_degree, closeness, betweenness, eigenvector, HITS_hubs, HITS_auths, Katz, PageRank, load centralities)

**Co-Linked Undirected Graph**: The graph was created with the most important nodes based on their centralities for Cloud Computing and Distributes Computing. The graph is a weighted undirected and connected graph with 612 nodes and 137915 edges, and has a graph density of 9.180 (density refers to the number of edges close to the maximal number of edges).

**Co-Linking Undirected Graph**: The graph (Gcin) is a weighted undirected and connected graph with 585 nodes and 82761 edges, and has a graph density of 0.077. In addition, the graph has 4 Louvain communities and modularity coefficient equal to 0.4572. Community "0" has 203 nodes, "1" has 184 nodes, "2" has 125 nodes and "3" has 100 nodes.

**Largest Strongly Connected Component**: The graph (GLCC) has 549 nodes and 27888 edges. It is a simple unweighted strongly/weakly directed graph with no isolates, not a tree and not a bipartite graph. It has a reciprocity of 0.5452.