Web and Social Computing (IT752)
Lab Assignment 4

Submitted to : Dr. Sowmya Kamath
Submitted by : Shaikh Sahil Ahmed
Roll number: 192IT022

Dataset 1: wiki-Vote
Dataset 2: p2p-Gnutella04
Dataset 3: p2p-Gnutella06

The above 3 datasets were used and I concluded various properties as:
1. Degree centrality
2. Closeness centrality
3. Betweeness centrality
4. Eigen vector centrality

I have calculated all the centrality for every dataset and stored it into results directory. I am just showing for few nodes of the graph.

1. Degree Centrality
   Degree centrality is defined as the number of links incident upon a node.
               Score C= Number of edges attached to node.
               Standardized Score = Score/(n-1)   where n is the number of nodes.
Few node id with there degree centrality is shown as **('node id': degree centrality)**

Dataset 1:
'30': 0.003935901040202418
'1412': 0.004076468934495361
'3352': 0.06705088557773405
'5254': 0.040905257239246554
'5543': 0.03696935619904414

Dataset 2:
'1': 0.0012871196101866323
'2': 0.000827434035119978
'3': 0.0014709938402132942
'4': 0.0014709938402132942
'5': 0.0006435598050933162

Dataset 3:
'1': 0.0004589261128958238
'2': 0.0005736576411197798
'3': 0.0013767783386874712
'4': 0.0006883891693437356
'5': 0.0011473152822395595

2. Closeness Centrality
   It defined as the reciprocal of the average shortest path length.
        Score = 1 / avg (L(n,m)) where L(n,m) is the length of the shortest path between two nodes n and m.
Few node id with there closeness centrality is shown as **('node id': Closeness centrality)**

Dataset 1:
'30': 0.3503264181902435
'1412': 0.33431826861934516
'3352': 0.4286618709380619
'5254': 0.4085203786616709
'5543': 0.40674420310227233

Dataset 2:
'1': 0.2411025205168517
'2': 0.23262233088184806
'3': 0.23932471314818762
'4': 0.2505299623905175
'5': 0.224291939842574

Dataset 3:
'1': 0.2176931914681053
'2': 0.2150028367744641
'3': 0.24222549537281493
'4': 0.252842886980738
'5': 0.23471118890534537

3. Betweeness Centrality

   For every pair of the network, count how many times a node can interrupt the shortest paths between the two nodes in the pair.
Few node id with there betweeness centrality is shown as **('node id': Betweeness centrality)**

Dataset 1:
'30': 0.00023102788733512074
'1412': 0.0015307038343196969
'3352': 0.009672471889382807
'5254': 0.008422351464718317
'5543': 0.00461356921601605

Dataset 2:
'1': 0.000586402262323544
'2': 0.00035982221697799177
'3': 0.0010039032318180656
'4': 0.0010647630374942833
'5': 0.0001679611308343717

Dataset 3:
'1': 3.7302005443322194e-05
'2': 0.00013664663724608848
'3': 0.0005910520413539953
'4': 0.00022647347628812495
'5': 0.00028723933675645075

4. Eigenvector Centrality

   It tries to generalize degree centrality by incorporating the importance of the neighbours (or incoming links in directed graphs). It computes the centrality of a node as a function of the centralities of its neighbours.
Few node id with there eigenvector centrality is shown as **('node id': Eigenvector centrality)**

Dataset 1:
'30': 0.003723099671145027
'1412': 0.0025628378029862505
'3352': 0.09178562117868826
'5254': 0.048006905578891944
'5543': 0.04762088963281069

Dataset 2:
'1': 0.004320704075988375
'2': 0.002361457165994293
'3': 0.00659597214997272

'4': 0.008445430184738513
'5': 0.0018904069445413666

Dataset 3:
'1': 0.0004336697977476668
'2': 0.00022377719307762662
'3': 0.001313538164089711
'4': 0.005470776221200113
'5': 0.0011626712595587635

PageRank Algorithm

The PageRank algorithm is applicable in web pages. Web page is a directed graph, we know that the two components of Directed graphs are -nodes and connections. The pages are nodes and hyperlinks are the connections, the connection between two nodes.

We can find out the importance of each page by the PageRank and it is accurate. The value of the PageRank is the probability will be between 0 and 1.

The PageRank value of individual node in a graph depends on the PageRank value of all the nodes which connect to it and those nodes are cyclically connected to the nodes whose ranking we want, we use converging iterative method for assigning values to PageRank.

I have calculated all the PageRank scores for every dataset and stored it into results directory. I am just showing for few nodes of the graph.

Hyperlink Induced Topic Search (HITS) Algorithm

Hyperlink Induced Topic Search (HITS) Algorithm is a Link Analysis Algorithm that rates webpages, developed by Jon Kleinberg. This algorithm is used to the web link-structures to discover and rank the webpages relevant for a particular search.

HITS uses hubs and authorities to define a recursive relationship between webpages. Before understanding the HITS Algorithm, we first need to know about Hubs and Authorities.

◆ Given a query to a Search Engine, the set of highly relevant web pages are called Roots. They are potential Authorities.

◆ Pages which are not very relevant but point to pages in the Root are called Hubs. Thus, an Authority is a page that many hubs link to whereas a Hub is a page that links to many authorities.

Dataset 1:

<u>**Part A:**</u>
PageRank scores for few nodes: **('node id':Score)**
'30': 0.00014746510313817384
'1412': 0.0003472198251112448
'3352': 0.0010023326587234476
'5254': 0.0010024806523846127
'5543': 0.0006768443933492331

HITS rank score for few nodes: **('node id':Score)**
Hub Score:
'30': 9.743151051640974e-05
'1412': 6.705823411581417e-05
'3352': 0.0024032860787707995

'5254': 0.001256864299457659
'5543': 0.0012468005099159677

Authorities Score:
'30': 9.743151051743071e-05
'1412': 6.705823411728649e-05
'3352': 0.0024032860787796297
'5254': 0.0012568642994453402
'5543': 0.0012468005098938907

Part B:

PageRank Min-Max Score: ('node id':Score)
min: '1971': 8.489877029144193e-05
**max:** '4037': 0.002601646674140776

HITS Min-Max Score: ('node id')
Min hub : ['2304', '2305', '3194', '3195', '3244', '3245', '4167', '4168', '4540', '4541', '5413', '5414', '5678', '5679', '5766', '5767', '5970', '5971', '6002', '6025', '6089', '6090', '6100', '6101', '6258', '6259', '6266', '6267', '7190', '7191', '7194', '7195', '7494', '7495', '7972', '7973', '7981', '7982', '8014', '8015']
**Max hub** : ['2565']
Min authority : ['2304', '2305', '3194', '3195', '3244', '3245', '4167', '4168', '4540', '4541', '5413', '5414', '5678', '5679', '5766', '5767', '5970', '5971', '6002', '6025', '6089', '6090', '6100', '6101', '6258', '6259', '6266', '6267', '7190', '7191', '7194', '7195', '7494', '7495', '7972', '7973', '7981', '7982', '8014', '8015']
Max authority : ['2565']

Plot for PageRank Score where score is plotted on y-axis and node-id is plotted on x-axis is shown in Fig. 1.
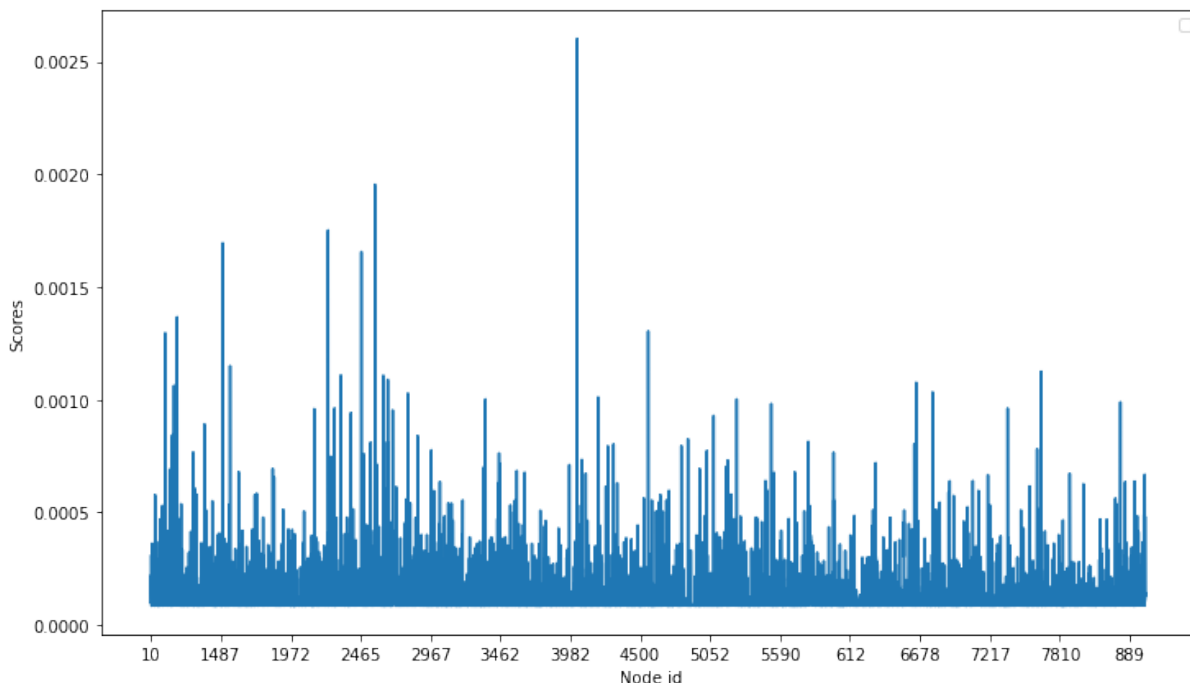
Fig. 1. Plot of PageRank for dataset 1.

Plot for HITS Score where score is plotted on y-axis and node-id is plotted on x-axis is shown in Fig. 2.
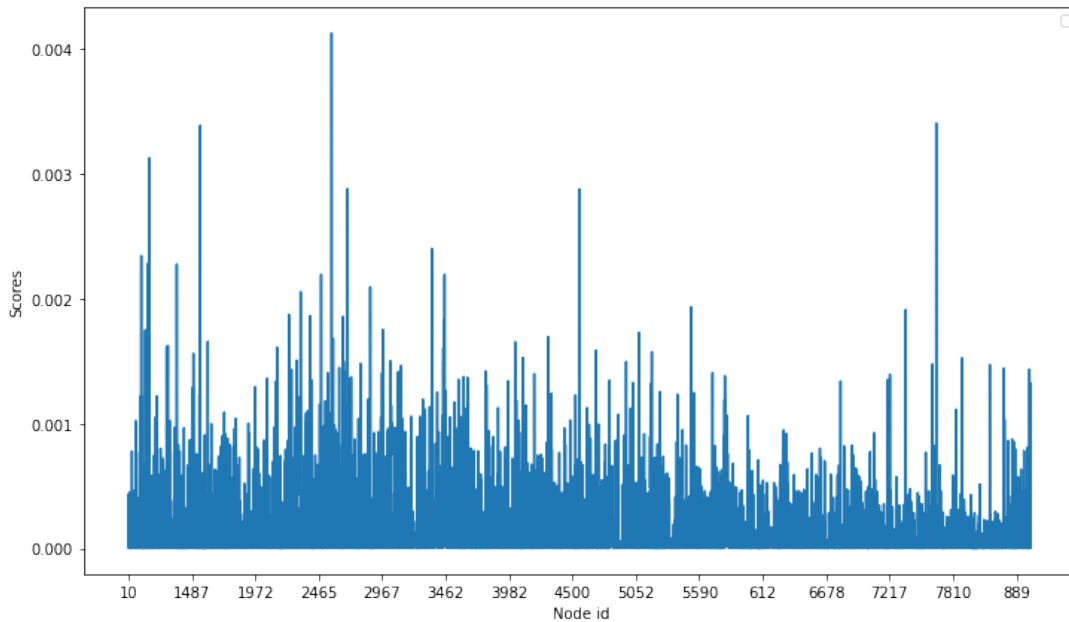
Fig. 2. Plot of HITS score for dataset 1.

Dataset 2:

Part A:
PageRank scores for few nodes: **('node id':Score)**
'1': 0.0001107277714508229
'2': 9.094474339932191e-05
'3': 0.00016187525129405137
'4': 0.0001202319191297852
'5': 8.336903514751971e-05

HITS rank score for few nodes: **('node id':Score)**
Hub Score:
'1': 9.492264524950195e-05
'2': 5.188087757389605e-05
'3': 0.00014493012764439257
'4': 0.0018555618350653367
'5': 4.153342130852476e-05

Authorities Score:
'1': 9.492264526487113e-05
'2': 5.188087757325573e-05
'3': 0.0001449301276339831
'4': 0.0018555618350463856
'5': 4.153342130852142e-05

Part B:
PageRank Min-Max Score: ('node id':Score)
min '2847': 5.693649003286417e-05
max '5598': 0.000750085171339918

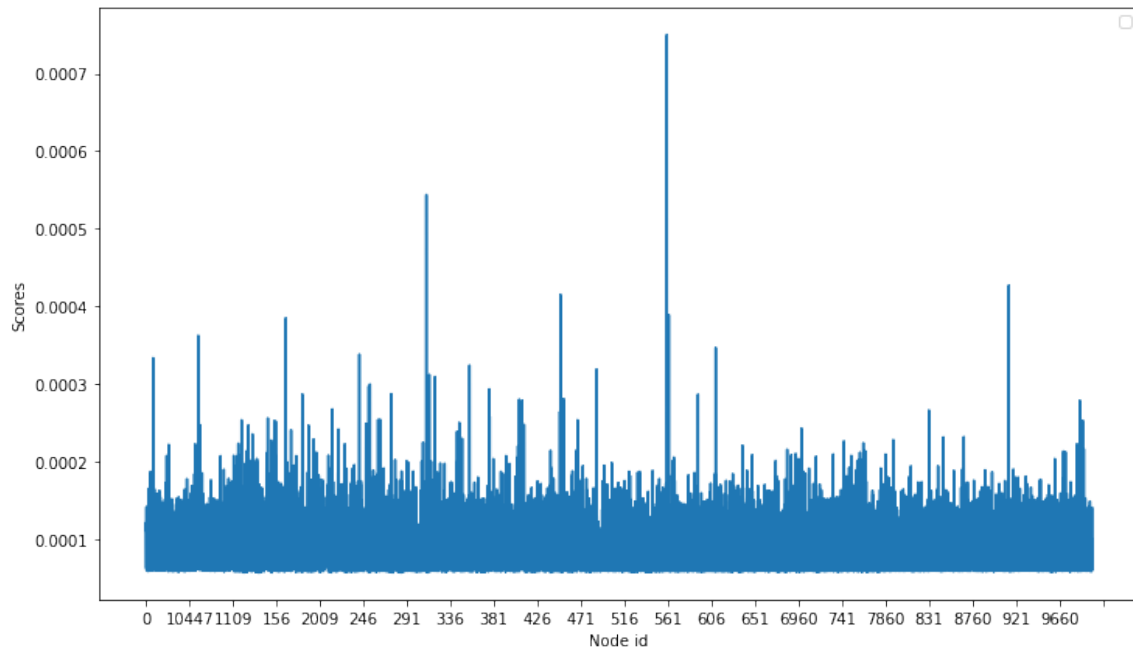**HITS Min-Max Score: ('node id')**
Min hub : ['src', 'dst']
Max hub : ['1054']
Min authority : ['src', 'dst']
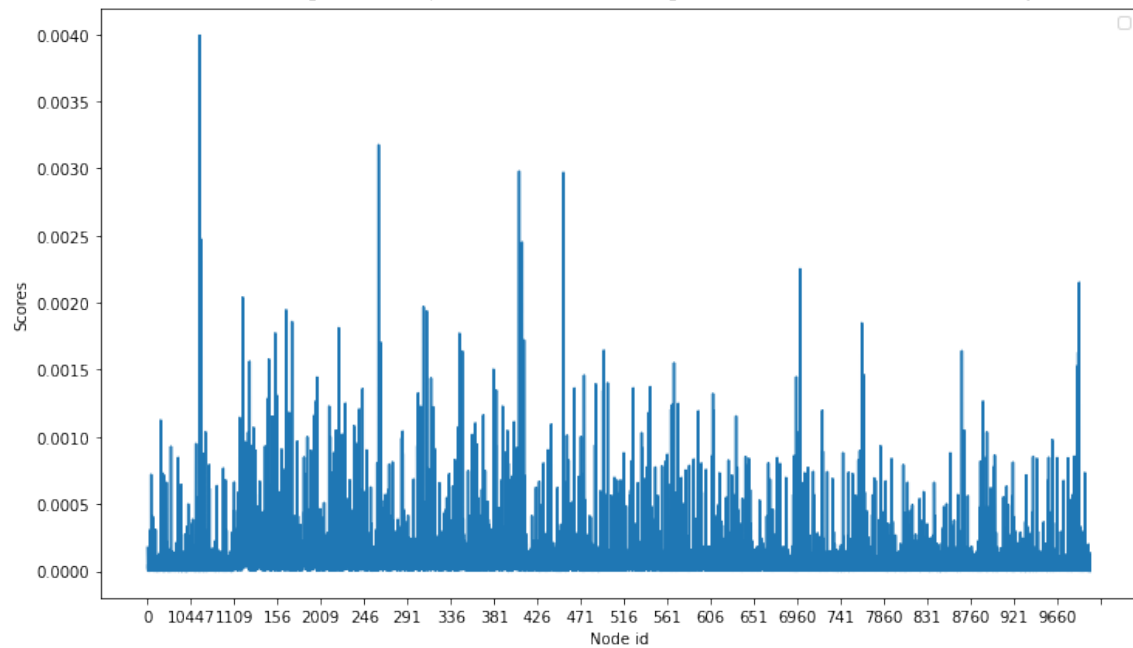**Max authority :** ['1054']

where 'src': 3.1295077523251376e-307, 'dst': 3.1295077523251376e-307

Plot for PageRank Score where score is plotted on y-axis and node-id is plotted on x-axis is shown in Fig. 3.

Fig. 3. Plot of PageRank for dataset 2.

Plot for HITS Score where score is plotted on y-axis and node-id is plotted on x-axis is shown in Fig. 4.

Fig. 4. Plot of HITS score for dataset 2.

Dataset 3:

Part A:

PageRank scores for few nodes: **('node id':Score)**
'1': 8.798679833698461e-05

'2': 0.0001072011792859982
'3': 0.0001278082395295808
'4': 9.019762783320757e-05
'5': 0.00011549218929176069

HITS rank score for few nodes: **('node id':Score)**
Hub Score:
'1': 1.6897733507375327e-05
'2': 8.733303789222012e-06
'3': 5.113602584846147e-05
'4': 0.0002130476109284547
'5': 4.530273317684226e-05

Authorities Score:
'1': 1.6897733507414754e-05
'2': 8.733303789130354e-06
'3': 5.113602584894833e-05
'4': 0.00021304761092982304
'5': 4.53027331769507e-05

Part B:

PageRank Min-Max Score: ('node id':Score)
min '2779': 7.081439410894645e-05
max '6494': 0.0008615517712639037

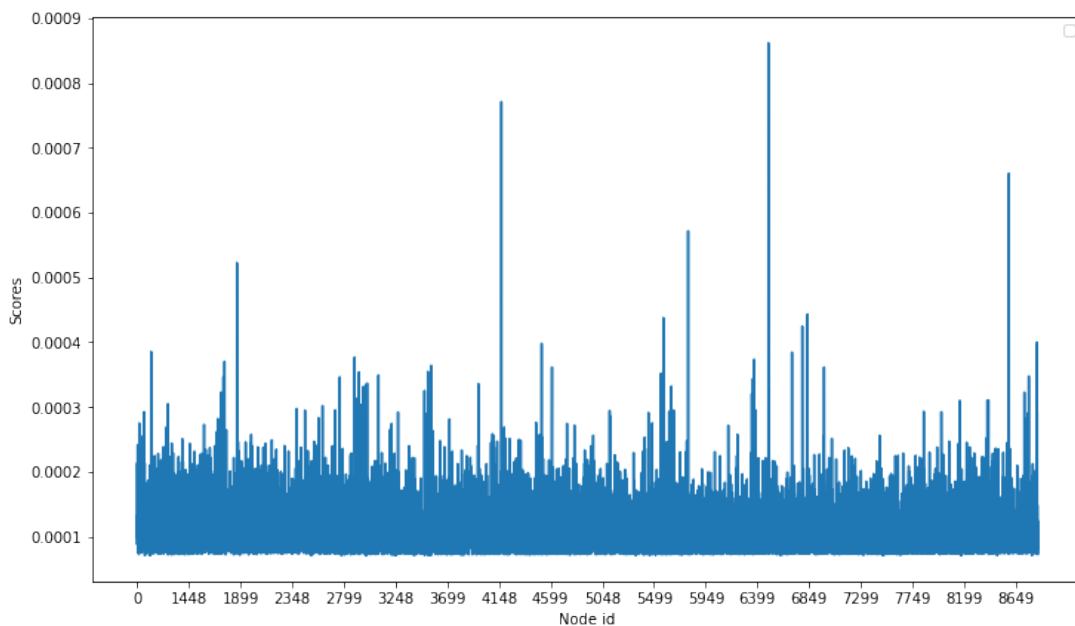### **HITS Min-Max Score: ('node id':Score)**
Min hub : ['7919']
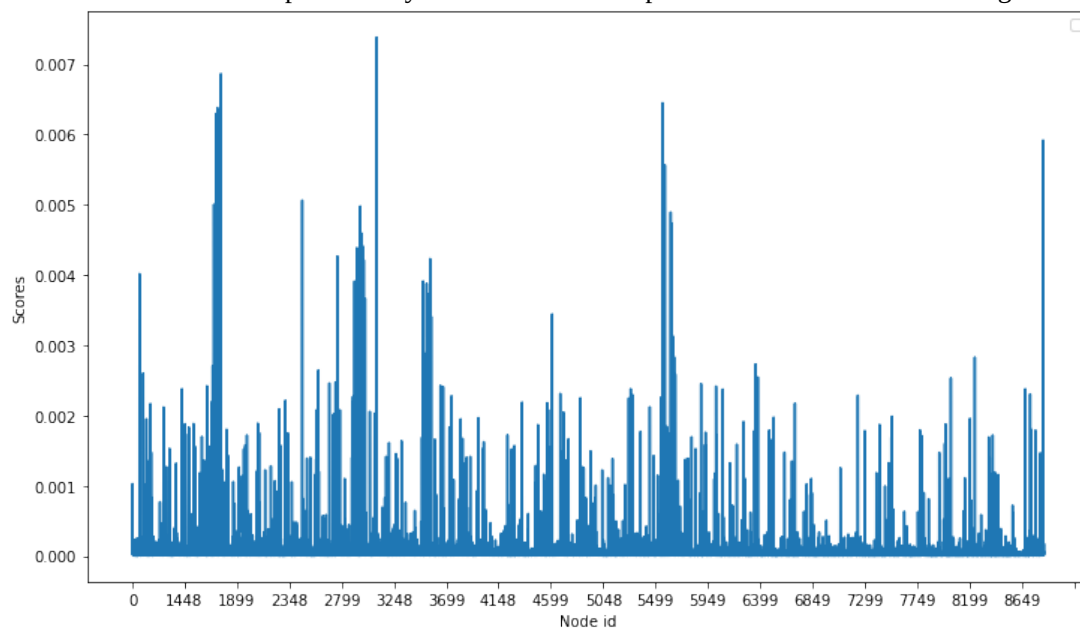**Max hub :** ['31']
Min authority : ['7919']
Max authority : ['31']

Plot for PageRank Score where score is plotted on y-axis and node-id is plotted on x-axis is shown in Fig. 5.

Fig. 5. Plot of PageRank for dataset 3.

Plot for HITS Score where score is plotted on y-axis and node-id is plotted on x-axis is shown in Fig. 6.

Fig. 6. Plot of HITS score for dataset 3.