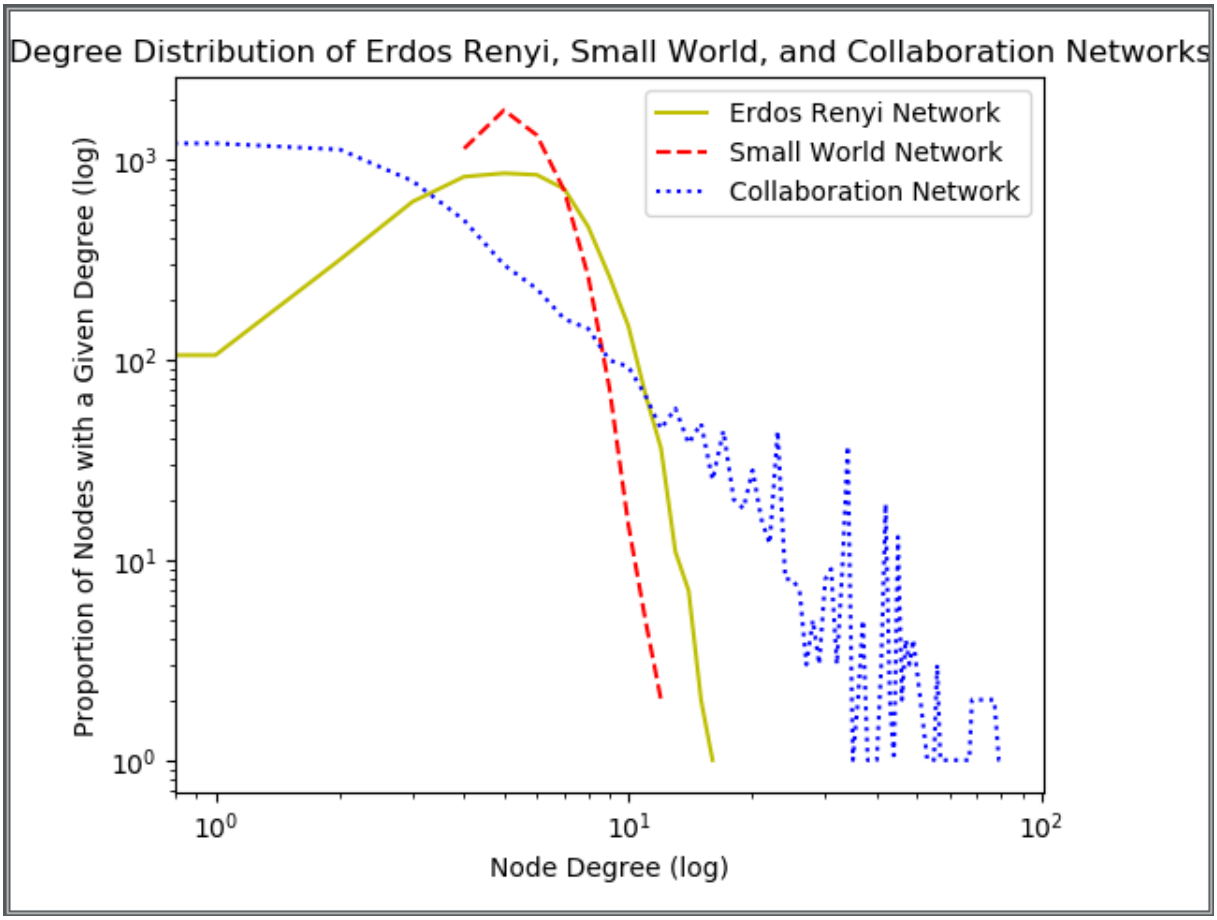
Homework 1

Mihael Trajbaric

1

1.1

Erdos Renyi Network has binomial degree distribution, as seen on picture below, while Collaboration Network has distribution, typical for real life network.



1.2

Average Clustering Coefficient for Erdos Renyi Network: 0.001616

Average Clustering Coefficient for Small World Network: 0.283884

Average Clustering Coefficient for Collaboration Network: 0.529636

The largest avg. clustering coefficient: Collaboration Netwrok

Comment on coeficiant: It makes sense that scientists usually work together in research groups and therefore publish together, therefore network has high Clustering.

2

2.1

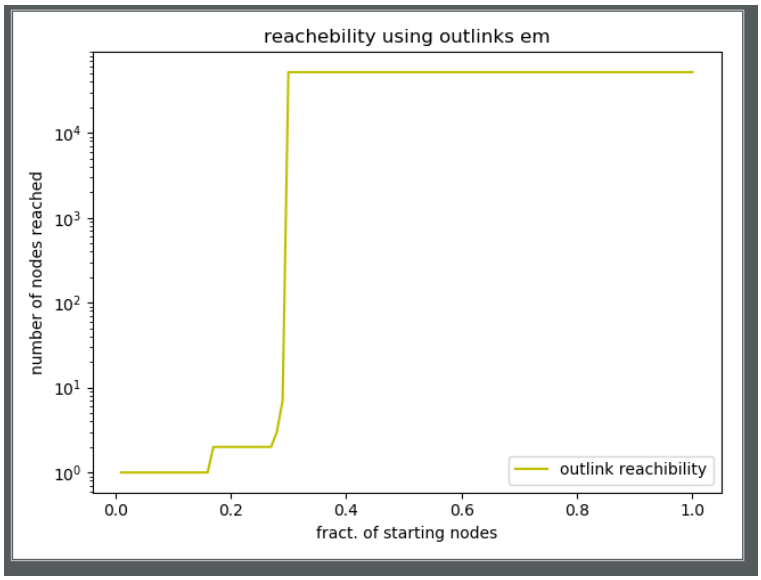
For determining whether node lies in SCC, IN or OUT we have several options. First option is that we pick random node from SCC, and then look if it is in a part of any of BFS trees. If it is part of outgoing tree, node is in IN, if it is a part of ingoing tree, node is in out and if it is part of both, node is in SCC

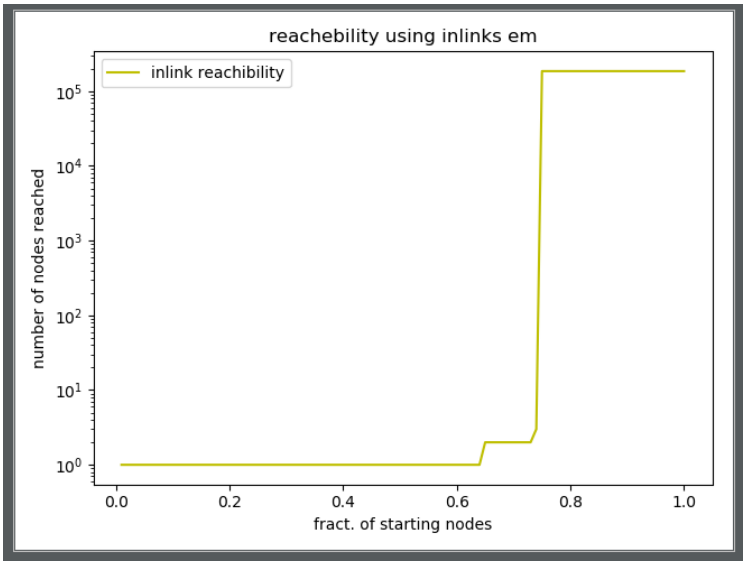
The second option is to compare sizes. Incoming tree with little nodes and outcoming tree with a lot means in, vice versa is OUT and if both trees are relatively big, node is located in SCC

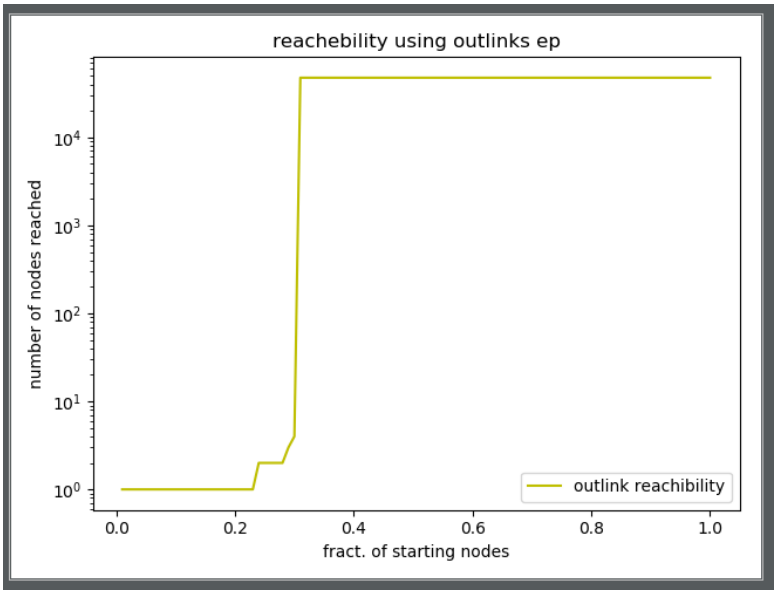
I used the first option. Node 2018 of email network is part of IN and node 224 of epinions network is part of SCC

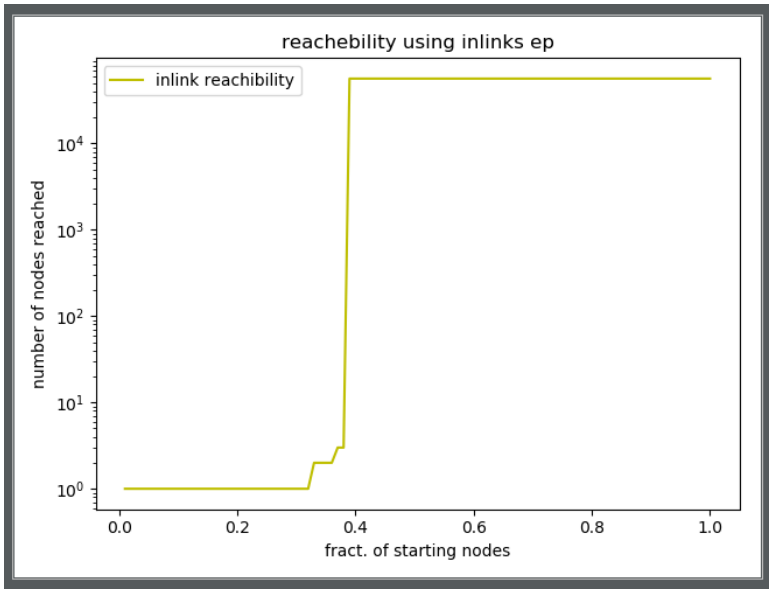
2.2

Comment: as for email network, it looks like IN is relatively huge, as seen from inlinks graph there is a pretty good probability of selecting node in this area. As for epinions network IN and OUT seems pretty close in size.









2.3

Size of the regions for email network: SCC 34203, IN 151024, OUT 17903, DISCONECTED: 40382, Tendrils and tubes: 21702

Epinions network: SCC: 32223, IN: 24239, OUT: 15457, DISCONECTED: 2, tendrils and tubes: 3958

SCC size is retrieved using snap's function. IN is calculated as #(nodes reached using in links bfs) – SCC size, OUT similarly to IN #(nodes reached using out links bfs) – SCC size, DISCONECTED components are calculated as all nodes minus size of WCC (weakly connected component) (all nodes - WCC) and TENDRILS_AND_TUBES = WCC - IN - OUT - SCC

2.4

Email probability = 0.731, epinions probability = 1.0

Expected probability as sample of nodes grow large should be getting close to ration #nodes in WCC/#all nodes. Nnodes in WCC can all be connected, while we cannot reach disconnected parts.

3

3.1

Yes, we can compute PPR score for Eloise as 3(A+C-B)-2*D. We cannot compute score for Felicity, since elements 4 and 5 always show up together, so it is not possible to compute PPR score with distinct values of 4 and 5. We can also compute PPR of Glynnis as (6A + 3B + 3C -2D)/10 (proofs in hard copy)

3.2

We can compute PPR score for every teleport set, which is linear combination of other sets.

3.3

S(b,c) = C and S(g,i) = C/2. In my opinion it does not make any sence, since (q,i) looks more similar.

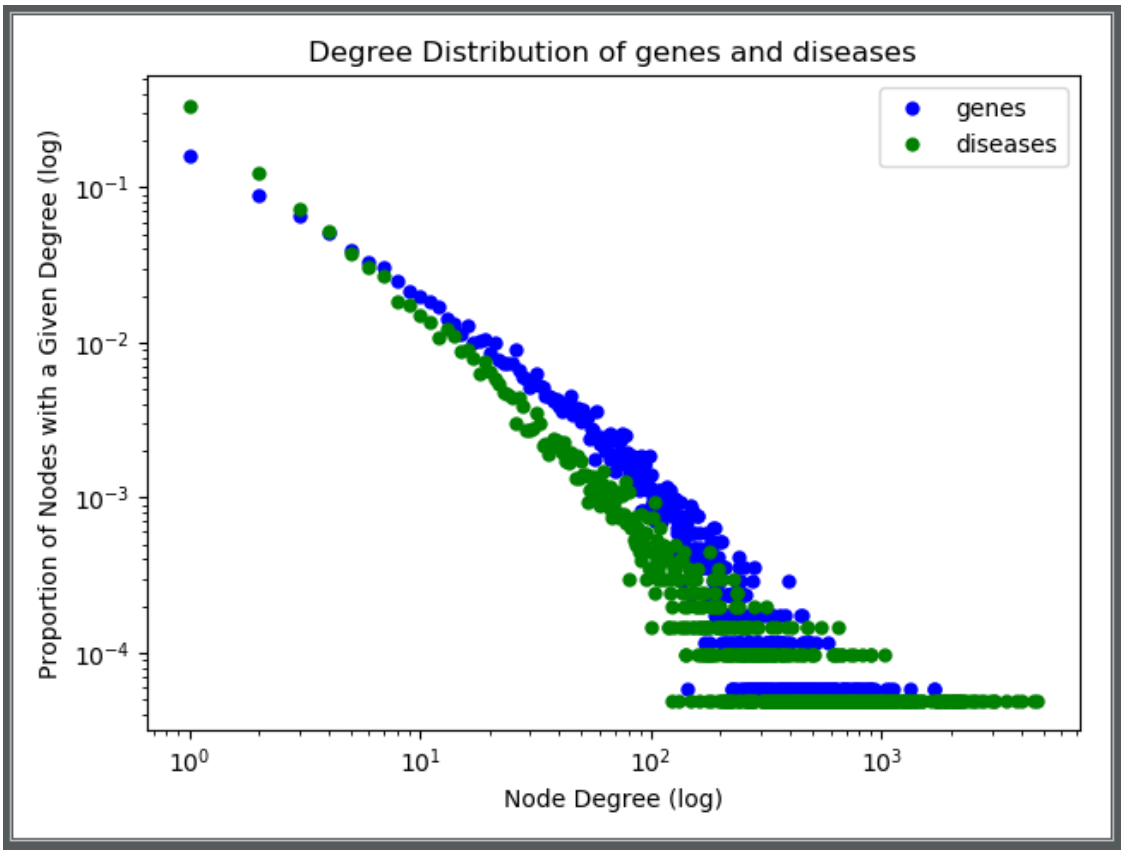
3.4

/

4

4.1

Disease network has 17074 genes and 20370 diseases. Total nodes: 37444, total edges: 561119.



Commentary on plot: Both have similar degree distribution, the horizontal lines of degrees of similar count for diseases are a bit lower than genes' because disease count is a bit larger.

4.2

Number of nodes of HDN network: 20370, number of edges: 12 892 536, density is 0.0621452030364, Clustering coefficient: 0.795414072556

4.3

The cliques arise because several diseases share same gene. The max clique size is therefore max degree among gene nodes. Therefore, calculating Kmax is simply calculating max degree, so it is linear.

Kmax = 1686

4.4

Contracting algorithm:

By inspecting degree count of initial graph (gene disease network) we can determine clique size. I have chosen to contract all the cliques of size greater than 250. For each clique (after identifying clique members by finding neighbours of gene with high degree in initial network) we can pick one as supernode, create a list of all the edges of all the nodes in clique, delete all the nodes in clique except the supernode, and then finally assign all the edges (except edges going to now deleted nodes of clique) to supernode. In case if one already contracted super node is part of another clique, we hold list of supernodes, and in case of finding one, we do not delete this node.

Scores: #of nodes: 9264, #edges 82382, density 0.001920048134750452, and C 0.7136209725783

4.5

Crohn’s Disease, CN

1. Malignant neoplasm of breast (score 549)
2. Breast 'Carcinoma (score 541)
3. Ulcerative Colitis (score 497)
4. Neoplasm 'Metastasis (score 493)
5. Rheumatoid Arthritis (score 492)

Crohn’s Disease, JA

1. Ulcerative Colitis (score 0.403409090909)
2. Inflammatory Bowel Diseases (score 0.367588932806)
3. Colitis (score 0.229166666667)
4. Rheumatoid Arthritis (score 0.218861209964)
5. Psoriasis (score 0.20563594821)

Leukemia, CN

1. Malignant neoplasm of breast (score 1361)
2. Breast Carcinoma (score 1346)
3. Carcinogenesis (score 1255)
4. Neoplasm Metastasis (score 1198)
5. Liver carcinoma (score 1133)

Leukemia, JA

1. Leukemia, Myelocytic Acute (score 0.417755572636)
2. Lymphoma (score 0.323611666001)
3. Acute lymphocytic leukemia (score 0.300361336947)
4. Multiple Myeloma (score 0.278726708075)
5. Melanoma (score 0.273531089561)

Commentary: as seen on list, the better metrics is Jaccard Index, because list of diseases is more similar. This is due to normalizing factor in index (when dividing by size of union) which prevents bias towards diseases with greater number of connected genes.