Web Mining Mini-Project Report

Jephy Rapheal
No: 15232756

Selection of dataset and Paper

The objective of the project is to study the structure of Online Social Network. To accomplish this I have selected YouTube **links** networks from the KONECT Network Collection as the dataset.

I have replicated the analysis and experiments described in the research paper "**Alan Mislove, Massimiliano Marcon, Krishna P. Gummadi, Peter Druschel, and Bobby Bhattacharjee. Measurement and analysis of online social networks. In*Proc. Internet Measurement Conf.*, 2007."** to implement an analysis on the network to confirm the power law, small-world and scale-free properties of online social networks. The Paper presents a large-scale measurement study and analysis of the structure of multiple online social networks like Youtube, Flickr, Orkut and Live Journal.

I have chosen **Youtube** dataset because of the following reasons

- Youtube is one among the most popular sites in the internet. This social network provides a powerful means of sharing and publishing contents and connecting with friends.
- The enormous datasets available for Youtube gives us an opportunity to analyse the structural properties of social networks at large scale.
- Youtube is one of the best example to analyse the impact of these networks on internet and people.

I have also used **Digg friends** networks from the KONECT Network Collection for testing the network properties.

Objective

The paper conducted an in-depth analysis of the graph structure of online social networks to evaluate the current online social systems. By trying to analyse and replicate the experiments conducted in the paper, I hope to get a good understanding of the structural properties of networks at large scale networks which includes:

- Link Symmetry
- Power-law node degrees
- Correlation of in-degree and out-degree
- Path lengths and diameter
- Link degree correlations
- Densely connected core

Understanding the graph structure of online social network is important to evaluate the current systems, to design future online network based systems and to understand the impact of these social networks.

Previous experiments/techniques

Earlier studies have revealed that:

- Average path length between two Americans in 6 hops.
- Social networks can be partitioned into strong and weak ties, and the strong ties are tightly clustered
- Social networks possess a large strongly connected component.
- Users in online social networks tend to form a tightly knit groups
- Social networks exhibits small-world behaviour as well as significant local clustering.

However previous systems have not yet been evaluated on real social networks at scale, and little is known on how to synthesize realistic social network graphs.

Statement of hypothesis

The paper intent to prove the hypothesis that online social networks have high degree of reciprocity (nodes with high in-degree also have high out-degree). Social network also appear to be composed of highly connected clusters consisting of relatively low-degree nodes. These clusters connect to each other via a small number of highly connected nodes. This implies clustering coefficient is inversely proportional to the node degree. The paper also intent to check if the network contain a large densely connected core and that almost all shortest paths in the network traverse through the highly connected core.

Data preparation steps

Dataset selected to carry out the replication of experiment is **Youtube links** networks from the KONECT Network Collection. This dataset includes social network of Youtube users and their connections. The network is directed and unweighted. Dataset provides list of **from** node in the first column and **to** node in the second column. There are 1,138,499 vertices (users) and 4,942,297 edges (links).

Datset selected for testing is Digg Friends Network. There are 279630 vertices (users) and 1731652 edges (links). Third and Fourth column in the dataset corresponds to Weight and timestamp of the edge. These columns are discarded as we do not require them for our analysis.

Web Mining Mini-Project Report

Jephy Rapheal
No: 15232756

Experimental settings, Results & Inferences

R was used to carry out the experiments as it can be conveniently used for statistical analysis and plotting. Within R I have used **igraph** package as it provides a wide range of functions that can be used for graph network analysis. Due to computational delays certain experiments were carried out on subset of datasets.

Detailed analysis and experiments which were carried out includes:

1) Link Symmetry

In directed graphs, there exists a symmetry if a connection from one node to another is reciprocated. It is important to check the symmetric nature of a network as increase in the symmetry increases the overall connectivity of the network and reduces its diameter. Increase in Symmetry can also make it hard to identify reputable source (nodes), because reputed sources tend to dilute their importance if they reciprocate back to any arbitrary users who link to them.

Code:

```
g=graph.data.frame(dat[,1:2],directed=TRUE)
#The proportion of reciprocated ties (for a directed network).
rec<-reciprocity(g,ignore.loops = TRUE)#Calculates the reciprocity of a directed graph.

#Alternate method
rec_calc<-2*dyad_census(g)$mut/ecount(g) # Calculating reciprocity.
```

Results:

Number of Users =1138494
Number of friends link =4942297
Fraction of Links symmetric =78.99202%

Inferences:

- The analysis of link symmetry of social networks with directed links reveals that they have a significant degree of symmetry.
- This implies that connectivity of the network is quite high.
- This also implies that we cannot easily identify the reputable source with the network as majority of connections are reciprocated back.

2) Power-law node degrees

Power law networks are networks where the probability that a node with degree k is proportional to k^-γ for large K and γ>1. The network which satisfies power law has the majority of nodes with small degree, and a few nodes with significantly higher degree.

To confirm if the online social networks are power law networks, the study examines the graph structure by considering the node degree distribution. To test how well the degree distributions are modelled by a power-law, the best power-law fit is calculated using the maximum likelihood method. Power-law coefficient along with the KolmogorovSmirnov goodness-of-fit metric was estimated.

Code:

```
#Log-log plot of of Complementary cumulative distribution functions

# In degree
yin=degree.distribution(g,cumulative = TRUE,mode = c( "in"))
plot(yin,log="xy",ylab = "P(Indegree>=x)", xlab = "Degree (log scale)", main = "Log-log plot of in-degree distribution")


#Out-degree
yout=degree.distribution(g,cumulative = TRUE,mode = c( "out"))
plot(yout,log="xy",ylab = "P(outdegree>=x)", xlab = "Degree (log scale)", main = "Log-log plot of Out-degree distribution")

#Power-law coefficient estimates (α) and corresponding Kolmogorov-Smirnov goodness-of-fit metrics (D)[have used plfit]
#The 'plfit' implementation also uses the maximum likelihood principle to determine alpha for a given xmin; When xmin is not given in advance, the algorithm will attempt to find itsoptimal value for which the p-value of a Kolmogorov-Smirnov test between the fitted distribution and the original sample is the largest. The function uses the method of Clauset, Shalizi and Newman to calculate the parameters of the fitted distribution.

#In-degree
d <- degree(g, mode="in")
fit1 <- power.law.fit(d, 1)#1 - The lower bound for fitting the power-law
cat("Alpha = ")
fit1$alpha#the exponent of the fitted power-law distribution.
cat("test statistic of a Kolmogorov-Smirnov test = ")
fit1$KS.stat#the test statistic of a Kolmogorov-Smirnov test that compares the fitted distribution with the input vector. Smaller scores denote better fit.

#Out-degree
d <- degree(g, mode="out")
fit2 <- power.law.fit(d, 1)#1 - The lower bound for fitting the power-law

cat("Alpha = ")
fit2$alpha#the exponent of the fitted power-law distribution.
```

Web Mining Mini-Project Report

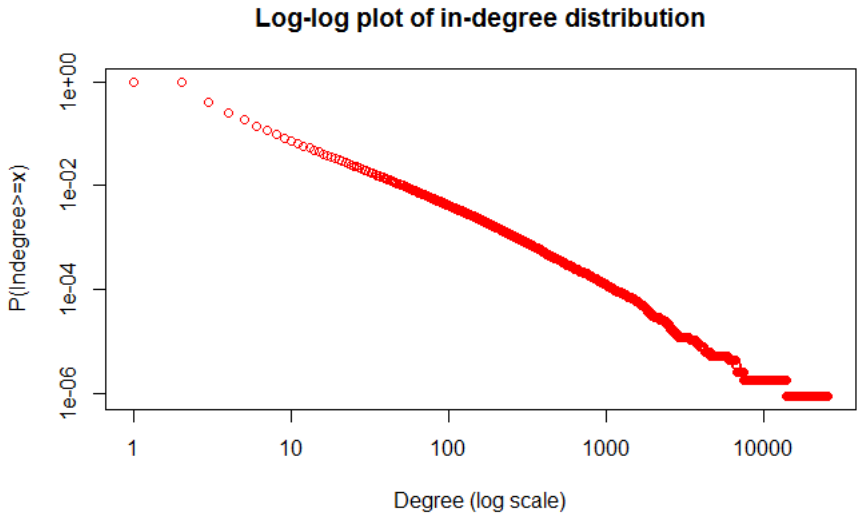
Jephy Rapheal
No: 15232756

cat("test statistic of a Kolmogorov-Smirnov test = ")
fit2$KS.stat#the test statistic of a Kolmogorov-Smirnov test that compares the fitted distribution with the input vector. Smaller scores denote better fit.

Results:

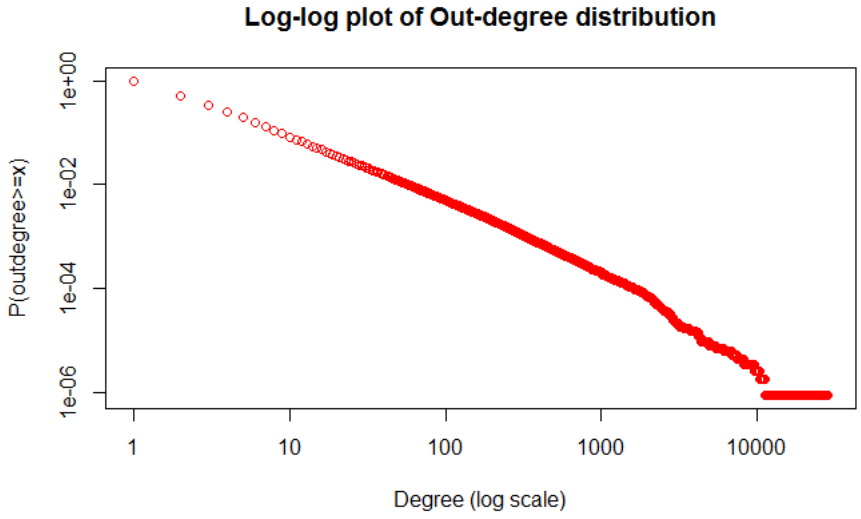
 # In Degree

Alpha = 1.985846
Test statistic of a Kolmogorov-Smirnov test = 0.008003207



 # Out Degree

Alpha = 2
Test statistic of a Kolmogorov-Smirnov test = 0.2864768



Inferences:

- It was found that the best power law coefficients approximate the distributions very well for Youtube networks.
- This implies that majority of nodes have small degrees, and only a few nodes have significantly higher degree.

3) Correlation of in-degree and out-degree

It is important to study the relation between in-degree and out-degree distribution in a web graph. In web, pages that are active (i.e have high out-degree) might not imply that they are very popular(i.e have high in-degree).

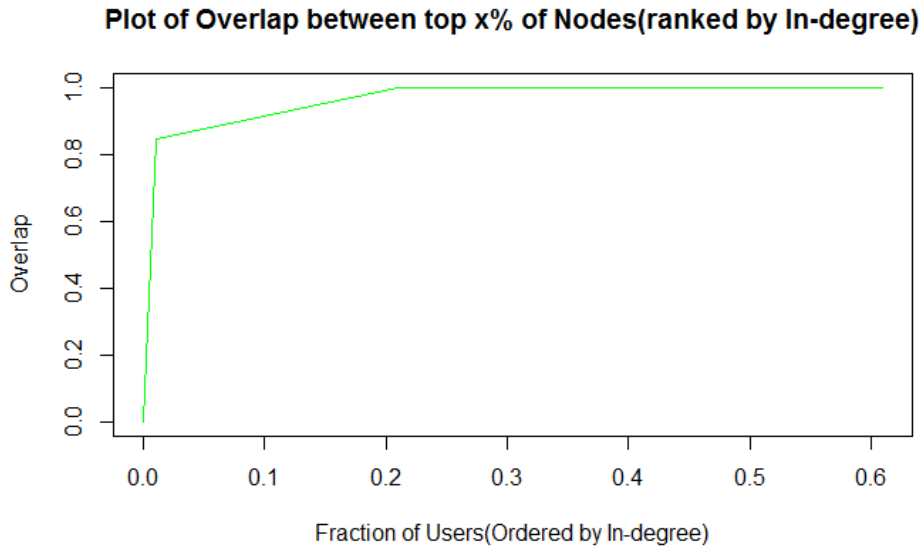
In social networks however, the nodes with very high outdegree also tend to have very high indegree. To confirm this, the study calculated the overlap between top x% of nodes ranked by outdegree and indegree.

Code:

```
#3. Corelation between indegree and outdegree

d.in <- degree(g, mode="in")
d.in<-sort(d.in,decreasing=TRUE)#sorting the degrees of nodes
d.out <- degree(g, mode="out")
d.out<-sort(d.out,decreasing=TRUE)#sorting the degrees of nodes

#calculation for x%
final<-NULL
fract_users<-seq(0.01,1.0,by=0.2)#x%
for(i in fract_users){
  d.in.top<-d.in[c(1:ceiling(i*length(d.in)))]
  d.out.top<-d.out[c(1:ceiling(i*length(d.out)))]
  count =0;
  for(n in d.in.top){
    if(is.element(n,d.out.top)==TRUE){
      count = count+1
    }
  }
#Calculate Overlap
  overlap1 = count/(length(d.in.top))
  final<-rbind(final,overlap1)

}
#Plot Fraction of Users (ordered by out/indegree) Vs Overlap
plot(x=fact,y=overlap,type="l",col = "green",ylab = "Overlap", xlab = "Fraction of Users(Ordered by In-degree)", main = "Plot of Overlap between top x% of Nodes(ranked by In-degree)")
```

Results:

Fraction of users	No of nodes	Overlap
1.00%	9630	0.8458498
21.00%	238855	0.9990422
41.00%	466554	0.9995094
61.00%	694253	0.9996703



Inferences:

- From the above results we can conclude that in social networks the nodes with very high out-degree also tends to have very high in-degree. From the graph it is clear that:
- 1% of nodes ranked by in degree has more than 84% of overlap with 1% of nodes ranked by out-degree. Curve tends to rise and reaches almost the maximum 20%. At 20% we can see that overlap reaches around 99% and flattens thereafter.

4) Average Path lengths and diameter

Average path length is estimated by finding the average number of steps along the Shortest path between all possible pairs of users in the network. In internet, a short average path length would facilitate a quick transfer of information at reduced costs.

While diameter is defined as the longest among the shortest path between any two nodes in the network. In other words, a graph's diameter is the largest number of vertices which must be traversed in order to travel from one vertex to another. A disconnected graph has infinite diameter. Hence shorter the diameter, more connected the graph would be.

Code:

```
#Average path length, radius and diamter
apl<-average.path.length(g, directed=TRUE)
rad<-radius(g, mode = c("all"))
diam<-diameter(g, directed = TRUE, unconnected = TRUE, weights = NULL)
```

Results:

Average Path Length - 5.104496
Radius – 13
Diameter - 21

Inference:

- From the above results we can see that average path length, radius and diameter of the network was significantly short in spite of having a huge size.
- Shorter average path length and diameter might be because of the high degree of reciprocity within the social networks.
- Short average path length would lead to the concept of a small world where every user is connected to every other user through a very short path.

5) Link Degree Correlations

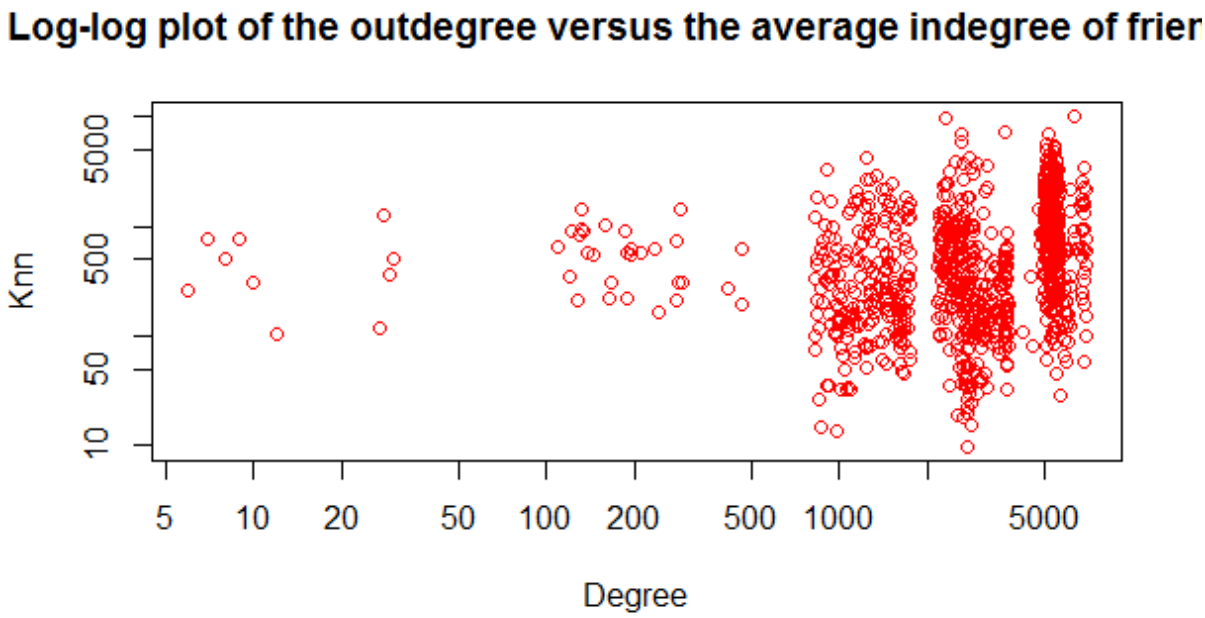
To further explore the differences in network structure between online social network and other networks, the paper did a study on how the users tend to connect to each other. Experiment was conducted to find the Joint degree distribution (JDD) to find how often nodes of a particular degree connect to each other. Basically we wanted to check if Nodes with high degree tends to connect more with nodes of high degree or not.

Experiment approximated a degree correlation function Knn, which is a mapping between out-degree of a node and the average in-degree of all the nodes incident to that particular node. Clearly, an increasing knn indicates a tendency of higher-degree nodes to connect to other high-degree nodes; a decreasing knn represents the opposite trend.

Code:

```
d.out.degree <- degree(g, mode="out")
d.out.degree<-sort(d.out.degree,decreasing=FALSE)#sorting the degrees of nodes

d.out.degree.sub<-head(d.out.degree,1000)
knn<-NULL
attr_vectr<-as.numeric(unlist(attributes(d.out.degree.sub)))
for(s in attr_vectr ){

  print(s)
  neigh_vector = neighbors(g, s, mode = "in")
  sum_1=0
  for(n in neigh_vector){
    sum_1 = sum_1 + degree(g,n, mode="in")
  }
  avg_indegree = sum_1/length(neigh_vector)
  knn<-rbind(knn,avg_indegree)

}
```

Results:



Inferences:

- From the above graph we can see that Knn value of Youtube network is decreasing, which shows that there is no trend for high-degree nodes to connect to other high degree nodes.
- This may be due its more “celebrity”-driven nature; that is there are a few extremely popular users on YouTube to whom many unpopular users connect.

Note: However other Social networks studied (Flickr, LiveJournal, Orkut) exhibited the trend of high-degree nodes connecting to other high-degree nodes, forming a core of network.

6) Densely connected core

Networks usually contain a densely connected core of high-degree and this core links small group of strongly clustered, low-degree nodes. A core must be necessary for the connectivity of the network (i.e., removing the core breaks the remainder of the nodes into many small, disconnected clusters). This can be checked by removing high degree nodes and analysing the connectivity of the remaining graph. We can then calculate the size of the largest remaining SCC, which is the largest set of users who can mutually reach each other.

Code:

```
#Densly Connected Core
d.high.degree <- degree(g, mode="all")
d.high.degree.sort<-sort(d.out.degree,decreasing=TRUE)
high.degree.index<-as.numeric(unlist(attributes(d.high.degree.sort)))
```

```
#
total<-length(d.high.degree)
fract<-c(0.01,0.1,1,10)
final<-NULL
for (f in fract){
densly.connected.core<-high.degree.index[c(1:(f*total)/100)]
v<-c(densly.connected.core)
sub_g<-induced_subgraph(g, -v)
clusters_weak = no.clusters(sub_g,mode=c("weak"))
clusters_strong = no.clusters(sub_g,mode=c("strong"))
cluster.num<-cbind(f,clusters_weak,clusters_strong)
final<-rbind(final,cluster.num)
}
```

Results:

Fraction of network removed(%)	Clusters_weak	Clusters_strong	Fraction of Clusters_weak	Fraction of Clusters_strong
0.01	2304	613779	**0.0037**	**0.9963**
0.1	7453	615684	**0.0120**	**0.9880**
1	39123	627442	**0.0587**	**0.9413**
10	189687	665506	**0.2218**	**0.7782**

Inferences:

- Above tables shows the Fraction of clusters (weak and strong) as we remove between 0.01% to 10% of the highest degree nodes.
- From the above results we can conclude that as we remove highest degree, the largest strongly connected components starts to split into smaller sized SCC

Testing

Digg friends networks from the KONECT Network Collection is used for testing the above network properties. Like youtube network this is again a directed friendship network and contains user links. A node represents a user, and a directed edge denotes that one user has the other user.

Below are the results and inferences after running the same steps on **Digg friends** Dataset:

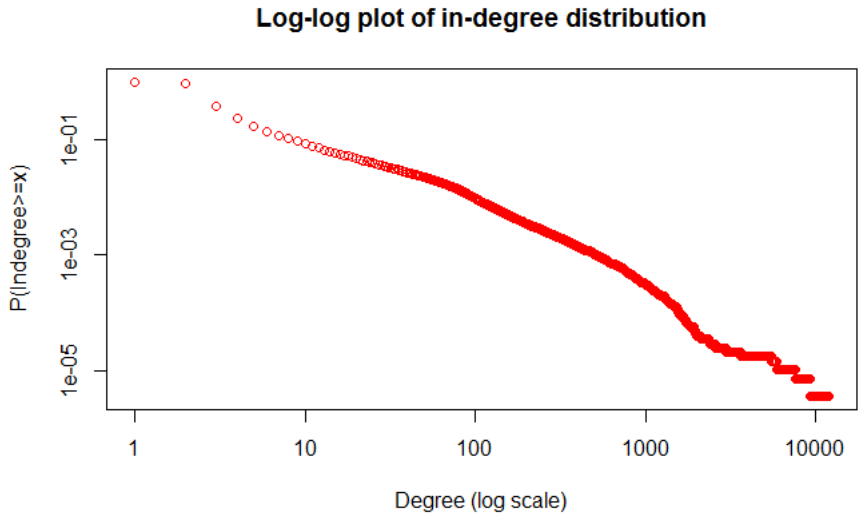
1.Link Symmetry

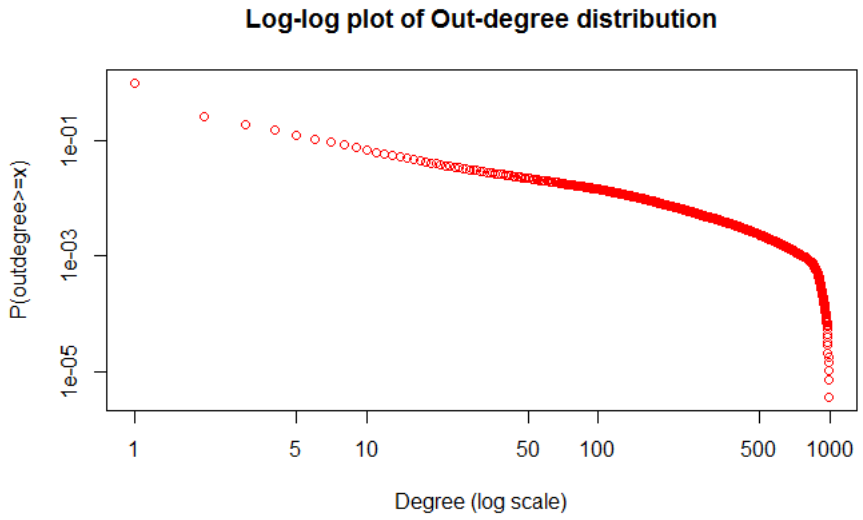
```
Number of Users =  279630
Number of friend link =  1731652
Fraction of Links symmetric =  21.19664 %
```

Inferences:

- The analysis of link symmetry of social networks with directed links reveals that they have quite a significant degree of symmetry.

2.Power Law Node degrees





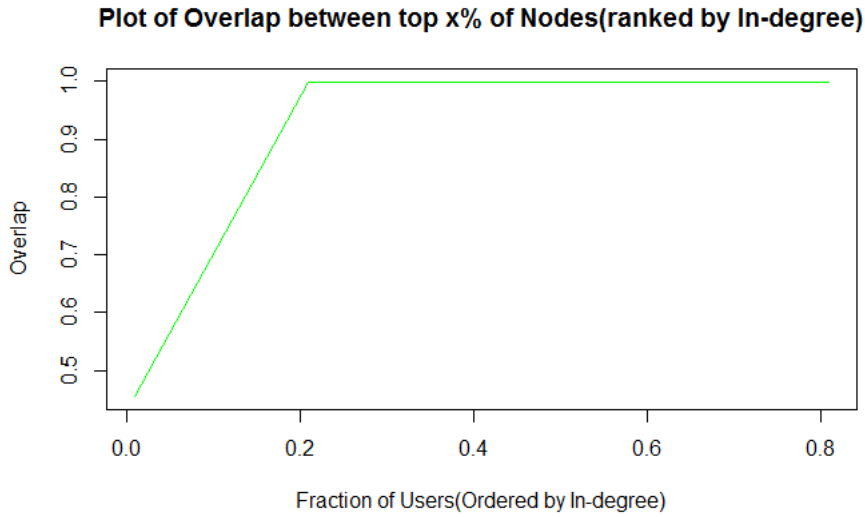

```
#In-degree
Alpha = 1.929941

test statistic of a Kolmogorov-Smirnov test = 0.01506258

#Out-degree
Alpha = 2
test statistic of a Kolmogorov-Smirnov test = 0.3555847
```

Inferences:

- It was found that the best power law coefficients approximate the distributions very well for even **Digg friends** networks.
- This implies that majority of nodes have small degrees, and only a few nodes have significantly higher degree.

3. Correlation between indegree and outdegree

Fraction of users	No of nodes	Overlap
1.00%	1273	0.4551305
21.00%	58575	0.9974797
41.00%	114501	0.9987091
61.00%	170427	0.9991323
81.00%	226353	0.9993466



Inferences:

- From the above results we can conclude that in social networks the nodes with very high out-degree also tends to have very high in-degree. From the graph it is clear that:
- 1% of nodes ranked by in degree has more than 45% of overlap with 1% of nodes ranked by out-degree. Curve tends to rise and reaches almost the maximum 20%. At 20% we can see that overlap reaches around 99% and flattens thereafter.

4. Average path length, radius and diameter

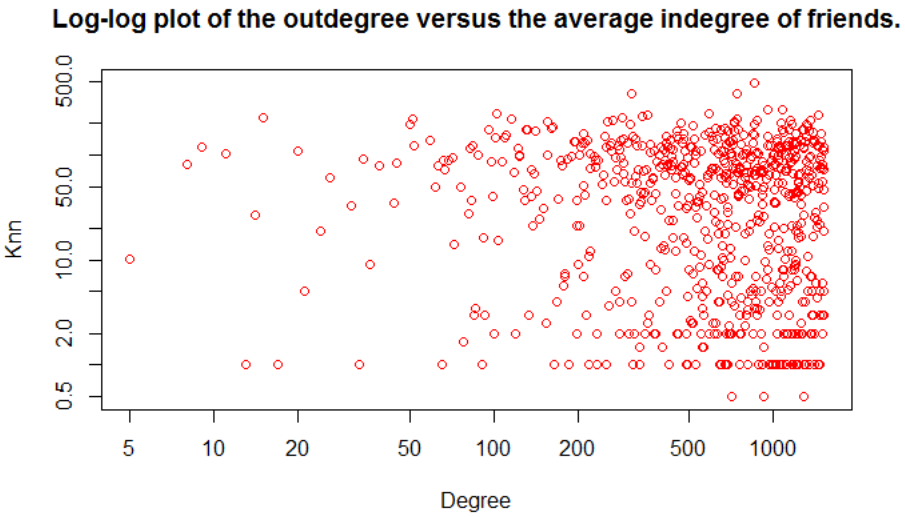
Average Path Length – 4.84023
Radius – 7
Diameter – 16

Inference:

- From the above results we can see that average path length, radius and diameter of the network was significantly short in spite of having a huge size.
- Short average path length would lead to the concept of a small world where every user is connected to every other user through a very short path.

5. Link Degree correlation

Joint Degree Distribution



Inferences:

- From the above graph we can see that Digg Friends network does not exhibit the trend of high-degree nodes connecting to other high-degree nodes, forming a core of network.

6. Densly Connected Core

Fraction of network removed	clusters_weak	clusters_strong	Fraction of clusters_weak	Fraction of clusters_strong
0.01	6908	240108	0.0279658	0.9720342
0.1	7425	240064	0.030001333	0.969998667
1	11951	239532	0.047522099	0.952477901
10	36679	224127	0.140637102	0.859362898

Inferences:

- Above tables shows the Fraction of clusters (weak and strong) as we remove between 0.01% to 10% of the highest degree nodes.
- From the above results we can observe that as we remove highest degree, th e largest strongly connected components starts to split into smaller sized SCC

Jephy Rapheal
No: 15232756

Conclusions

The experiments and analysis of the structural properties of youtube links and Digg Friends Networks reveals that social networks are structurally different from previously studied networks, in particular the Web. Below are some conclusions drawn:

- Social networks exhibits much higher link- symmetry when compare to other networks.
- Online Social networks also follows the power law.
- Online social networks also have high degree of reciprocity (nodes with high in-degree also have high out-degree).
- Social network appear to be composed of highly connected clusters consisting of relatively low-degree nodes. These clusters connect to each other via a small number of highly connected nodes.
- Clustering coefficient is inversely proportional to the node degree.
- Networks contain a large densely connected core and almost all shortest paths in the network traverse through the highly connected core.
- Both Youtube network and Digg Friends Network does not show any trend for high-degree nodes to connect to other high degree nodes(because of it “celebrity”-driven nature).