Analysis of Slashdot and Stanford Networks using Gephi and Cytoscape .

THESIS · DECEMBER 2015

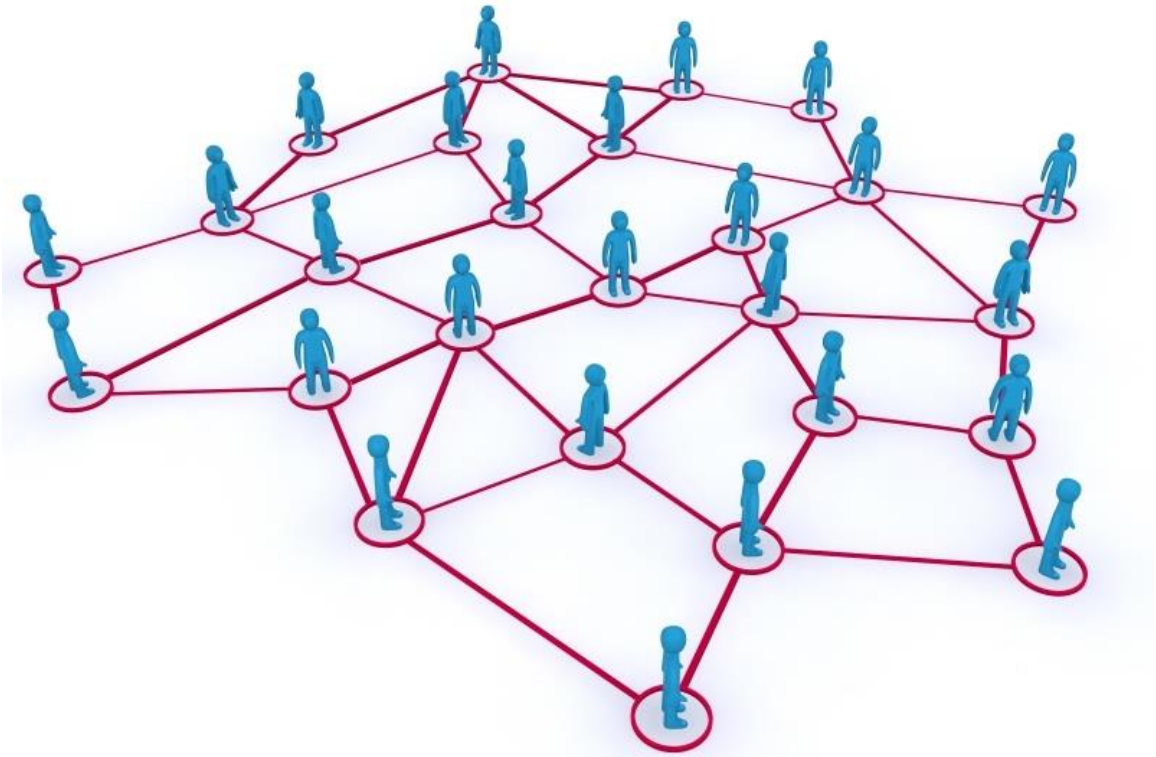
1 AUTHOR:

Ashutosh Agarwal

Shiv Nadar University

1 PUBLICATION **0** CITATIONS

SEE PROFILE



Analysis of Slashdot and Stanford Pages' Networks.

BY -ASHUTOSH AGARWAL

INTRODUCTION

The first network assigned to me is **Slashdot,** which is a news website that features news stories on Science and technology that are submitted and verified by its users. Each story has a comments section, where user can always post their views. A very interesting feature used in Slashdot is the ability to tag each of the user as a *"friend"* or a *"foe"*. Thus we have used this feature effectively to build a network, consisting of users as the **nodes** and the relation between them (friend or foe), is represented by an edge. There were *3348* nodes in total, with exactly *4622* edges. In other words, *3348* people on Slashdot were considered for the Network and the friend/foe relationships between them amounted to be *4622*.

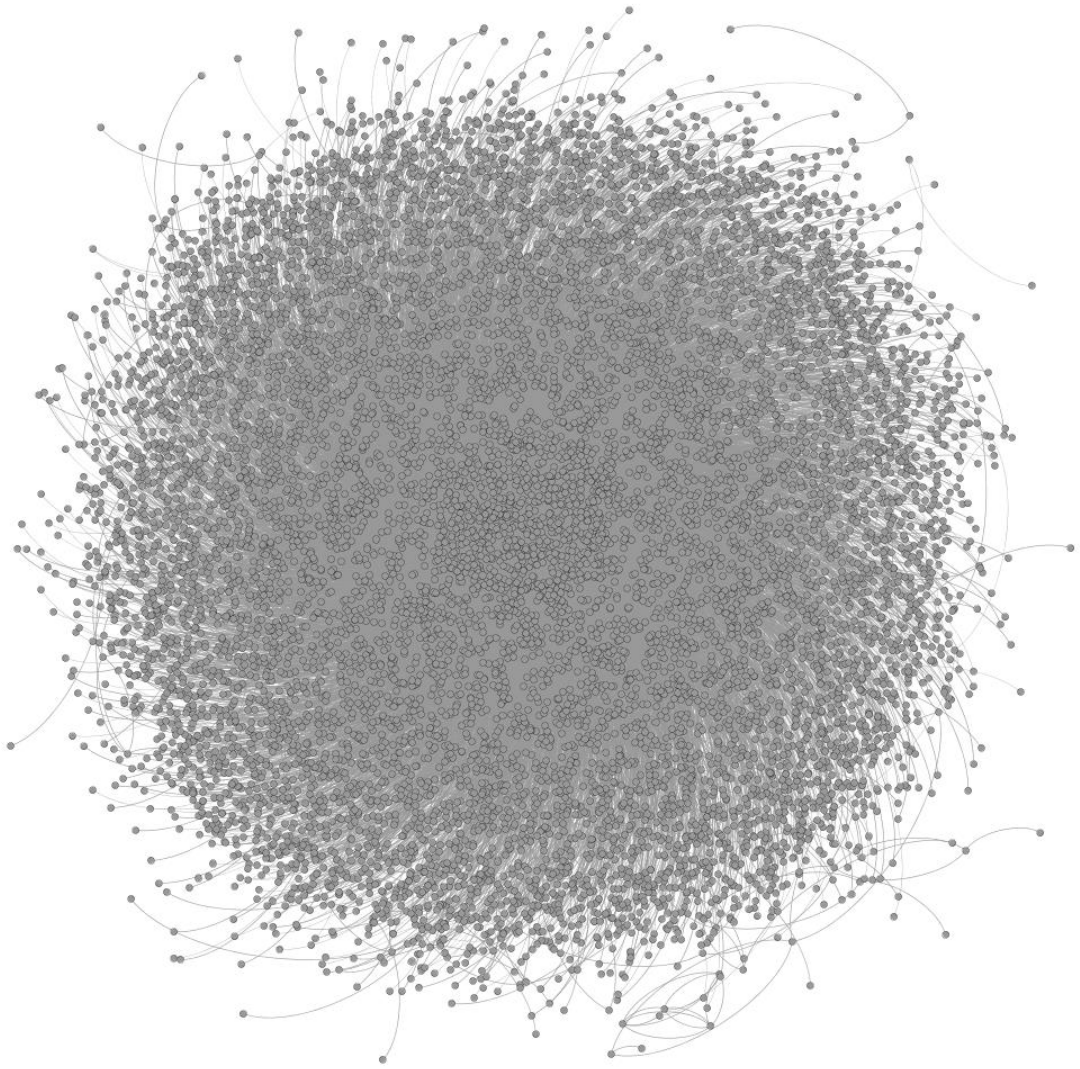
The second network assigned to me, is a web-graph which consists of web based pages from the ***Stanford University***. The pages have been considered as the nodes for this network. A hyperlink from one webpage to another represents an edge from node A to node B. There were *2401* nodes in total, with exactly *3603* edges. In other words, 2401 webpages were considered for the Network and the hyperlinks(directed) between them amounted to be 3603.

An interesting point, is that, this was a humongous Network. *Gephi* hanged a lot given about 2 billion nodes, and my processor is I3. It took a lot of time for the computation of this network, although, computation for Slashdot network was relatively easier. The two networks in themselves were quite different from each other. Slashdot had lesser no of edges(interconnections), while Stanford has more because having a friend/foe is less likely to happen than a page having links to other pages. Thus, it was found out that Stanford Network has more Average degree in comparison to Slashdot, which is a great find. I would like to make it clear that I have used Cytoscape for Network Analyzing owing to its great capability of finding the data in quick time.

VISUALIZATION

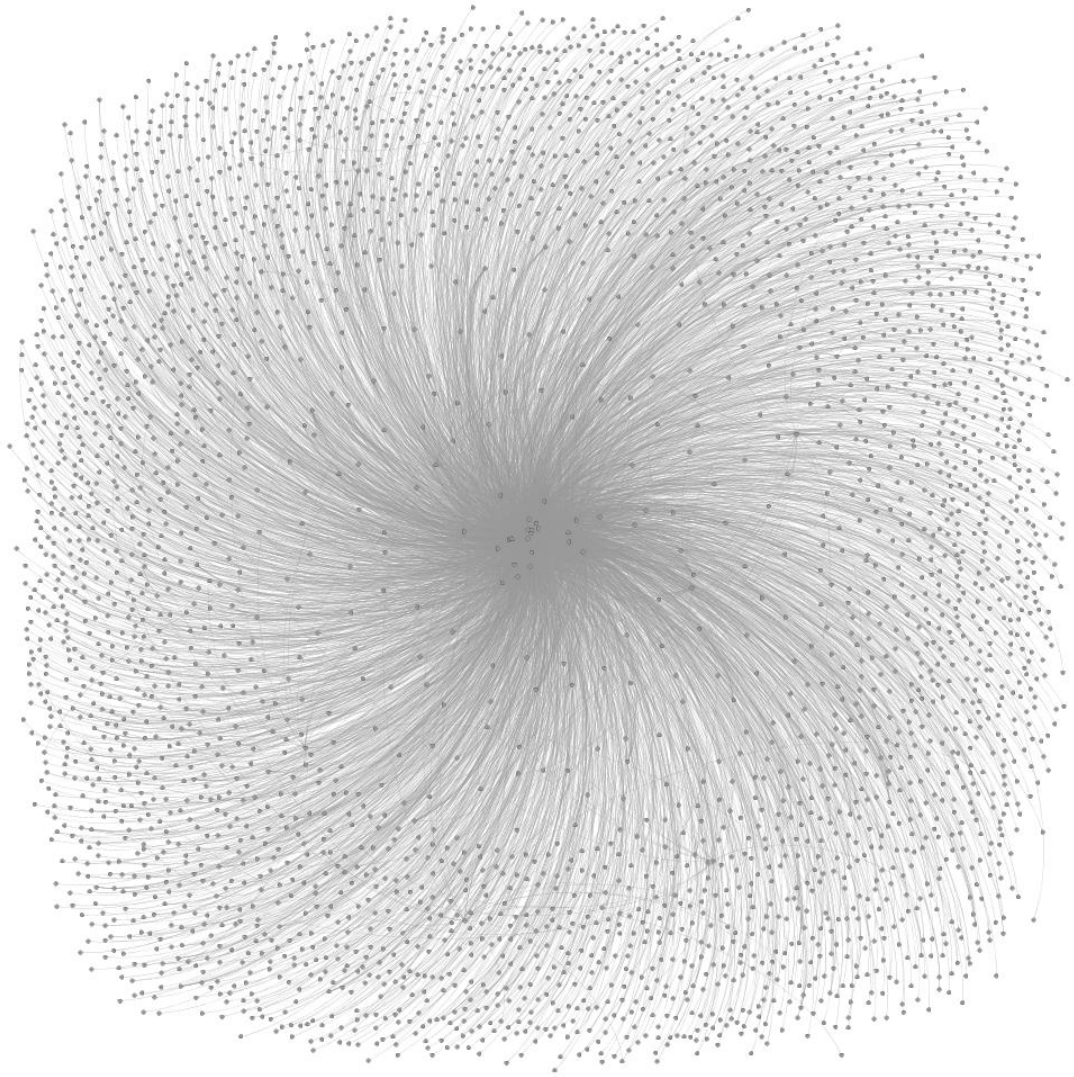
The following Graphs have been developed using the *Gephi* tool. Gephi uses Java as its compiler, and builds graphs based on the dataset we feed in mostly in CSV format.

The following graphs were saved in to the .png file format and were added to this Research paper, from there, as instructed to be done by the instructor.



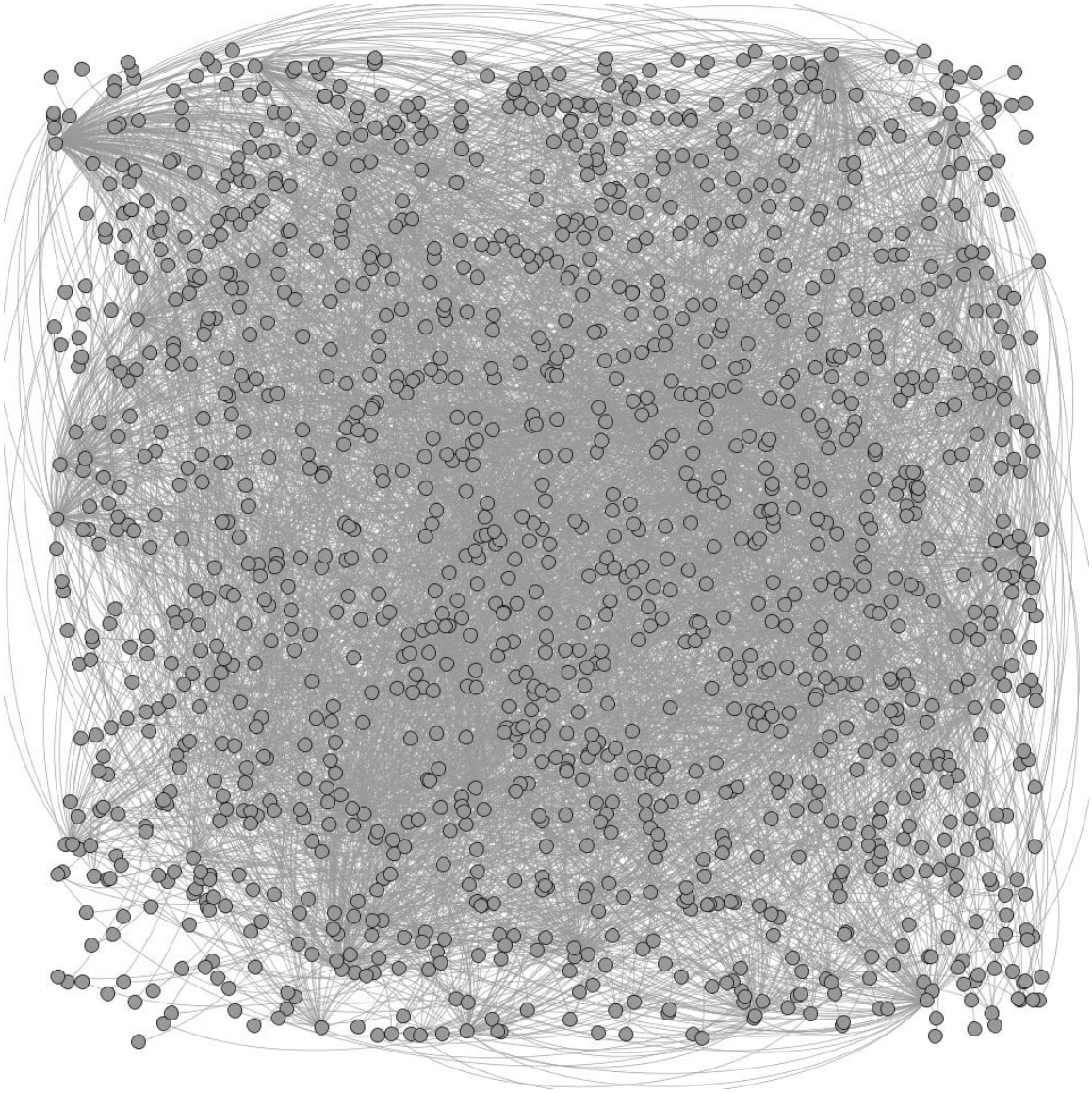
graph Slashdot file generated using the *Gephi* tool for Network Analysis.

Clearly many of the nodes are interconnected, with lot of interconnections in the middle, indicating a high clustering.



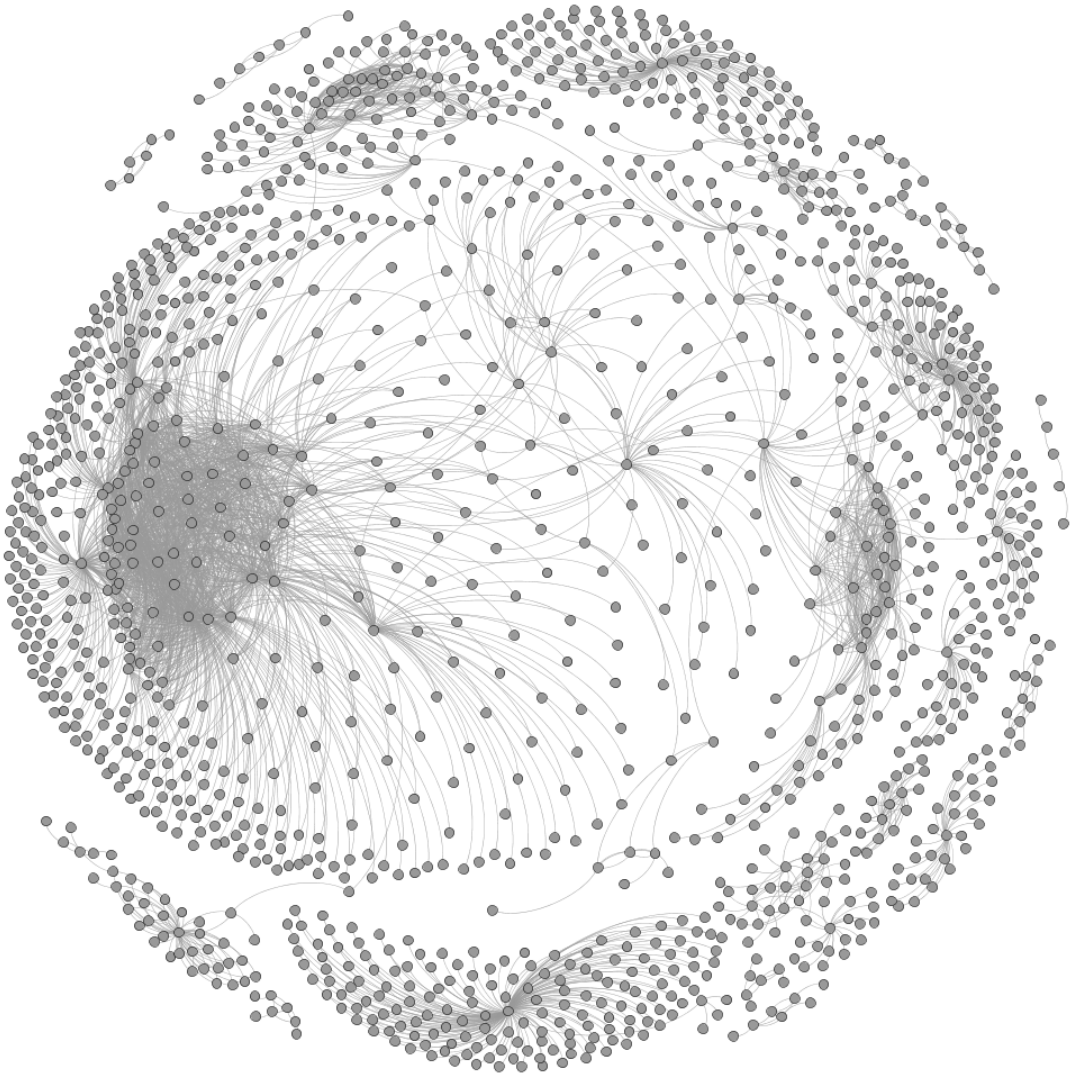
Connection of Edges in Slashdot Network (Fruchterman Reingold View)

Fruchterman Reingold View as Wikipedia says, is a way to present the graph in a pleasing way. Edges are made to be of more or less of equal lengths. This is specifically done to present graph from a detailed view, and clearly there are so many edges.



graph_Stanford file generated using the *Gephi* tool for Network Analysis.

Most of the nodes are scattered, implying low clustering as evident from data. Although, no of edges are significantly high.



Connection of Edges in Stanford Network (Fruchterman Reingold View)

Fruchterman Reingold View implying that nodes that are actually close together have more interconnections in comparison to those that are far apart. This is evident from the Average number of Neighbors data.

ANALYSIS

- **Random Graphs** are totally dependent on Probability distribution. This probability distribution may be done in any random manner. I can add edges between any vertices, randomly, so as to generate a graph. Thus these networks *provide a good intersection between graph theory and probability theory*. The aim of such a network would be to study or predict when a certain property of graphs would come in to picture. The most common of such random graphs is the $G(n, p)$, in which every possible edge occurs independently with probability $0 < p < 1$.
- A **scale-free network** is a one whose degree distribution would follow power law. The power law is a relationship between 2 quantities, where some change in one quantity would cause some change in the other quantity proportionally. The *networks used in this research paper are Scale-free*. Now, this means that in a scale free network a fraction of P(m) nodes in the network having m connections to other nodes goes for large values of m as

$$P(m) \sim m^{-\gamma}$$

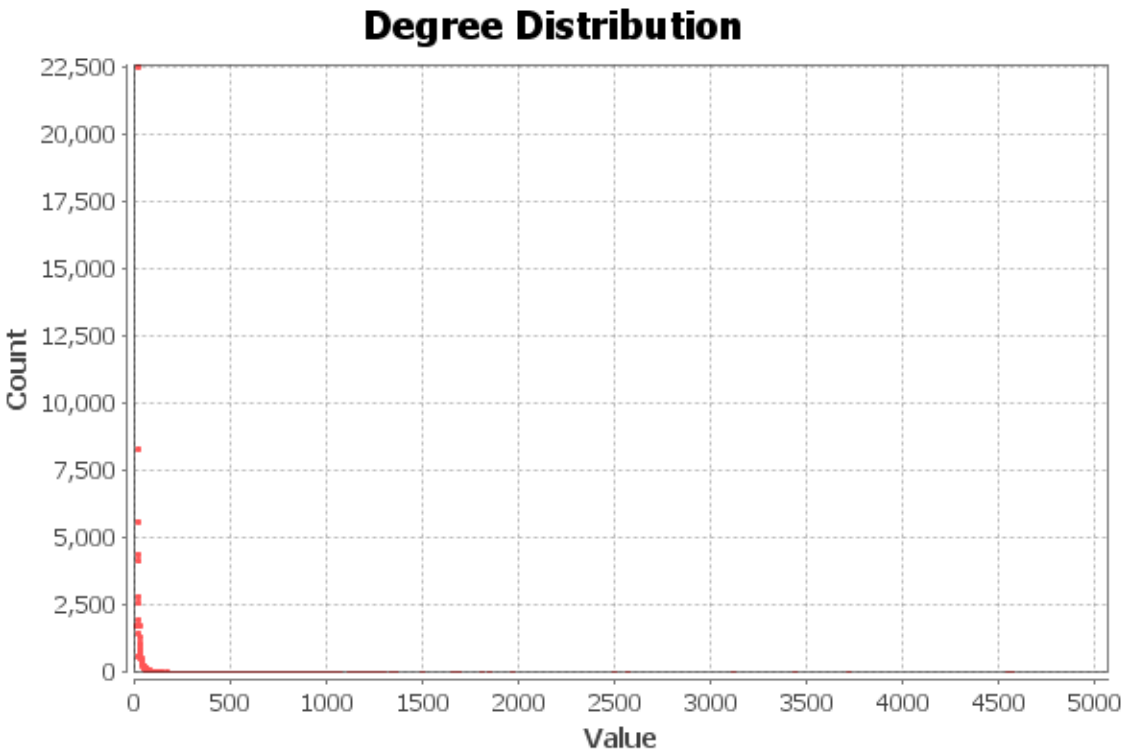
- Many Scale-Free Network Models have been proposed, such as, The Barabasi-Albert Model, Two-Level Network Model, Non-Linear Network Model etc.

- The **Barabasi-Albert Model** is an Algorithm for *generating Scale-Free Networks using a preferential attachment mechanism.* The network starts with an initial connected network of m_o nodes. After that we keep on adding newer nodes one at a time. Each new node is connected to m<= mo existing nodes with a probability that is in proportion to number of links that the existing nodes already have.

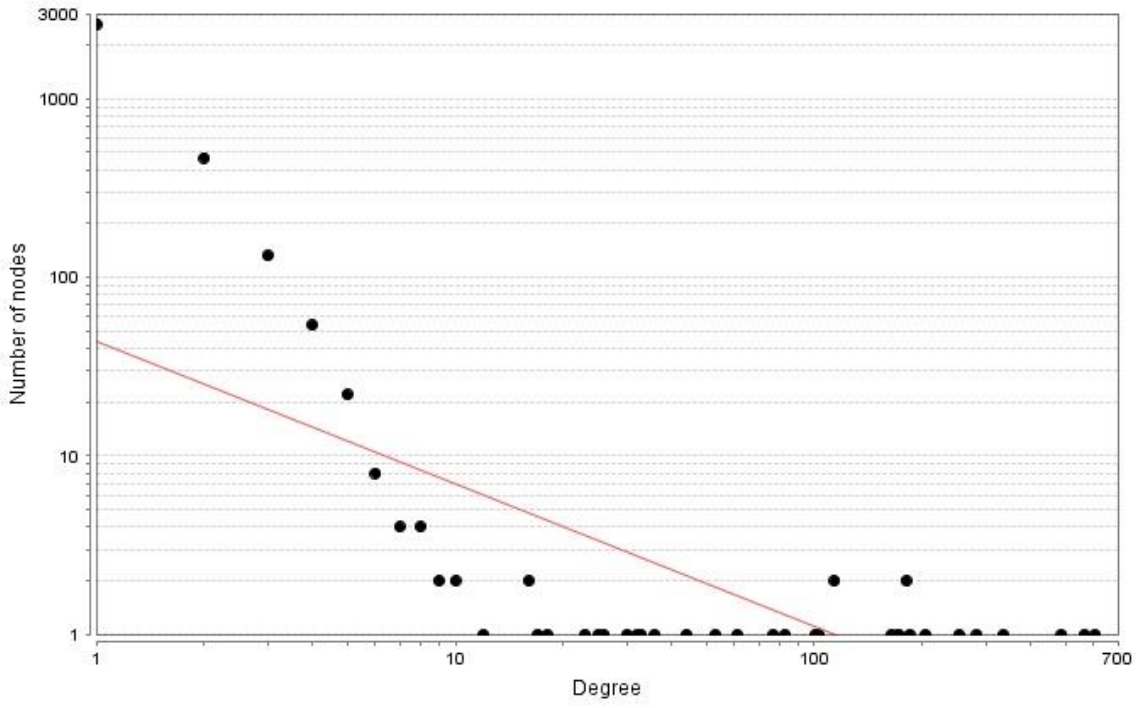
  The degree distribution resulting from the Barabasi-Albert model is scale free, thus establishing the power law in the form of

$$P(m) \sim m^{-3}$$

Slashdot Analysis



It is self-explanatory from the graph, that this is a classic example of the Power law. A lot of things can be inferred from the graph including the Pareto (80-20) principle. To the right is the long tail and to the left are the few that dominate.



Power Law fit using ***"Cytoscape"***

The data for Power Law was Calculated as follows in Cytoscape.

$y = ax^b$

a	43.815
b	**-2.031**
Corelation	0.834
R-Squared	0.519

DATA ON Slashdot

Average Degree	**1.381**
Average Clustering Coefficient	**0.095**
Average Path Length	**2.71**
Network Diameter	**4**
Avg. no. of Neighbors	**2.696**
Network Heterogeneity	**7.542**

Average Degree in this context implies that there are about **1.381** edges for one node in the Slashdot network, or more textually that implies ***that each person on Slashdot has about 1 friends or foes***.

Clustering Coefficient is the ability of nodes in a graph to cluster together with neighbors.

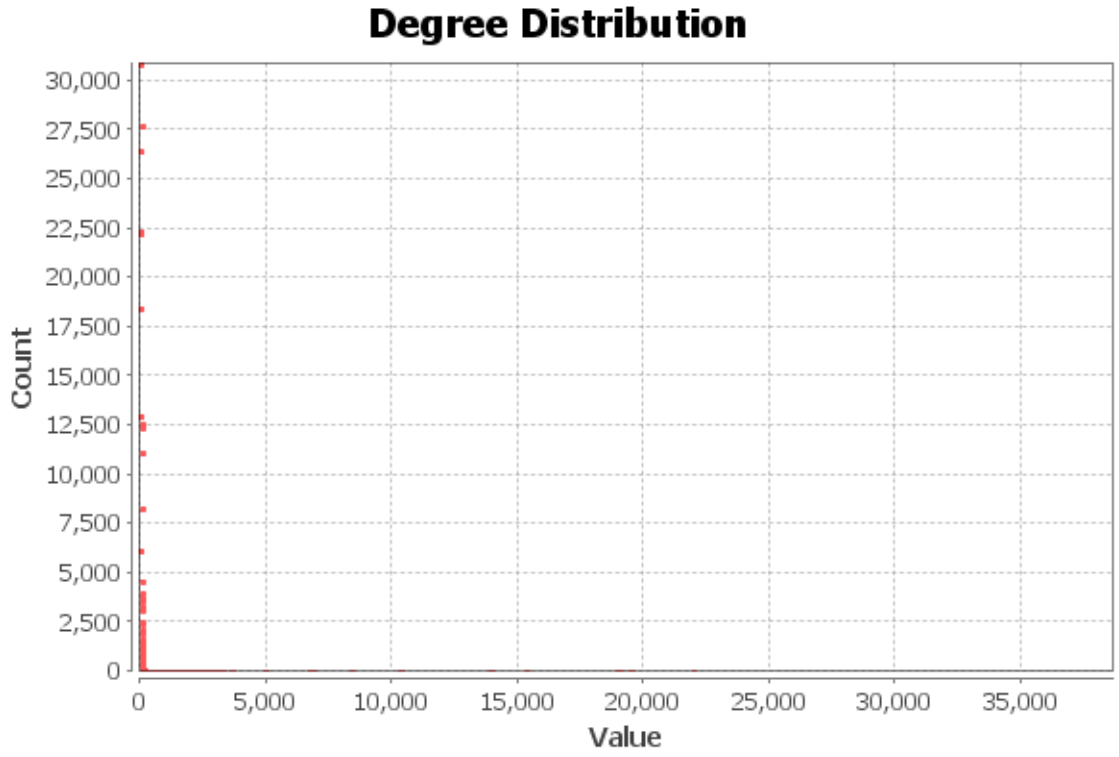
Average Path Length is the average number of steps along the shortest paths for all possible pair of nodes in the network.

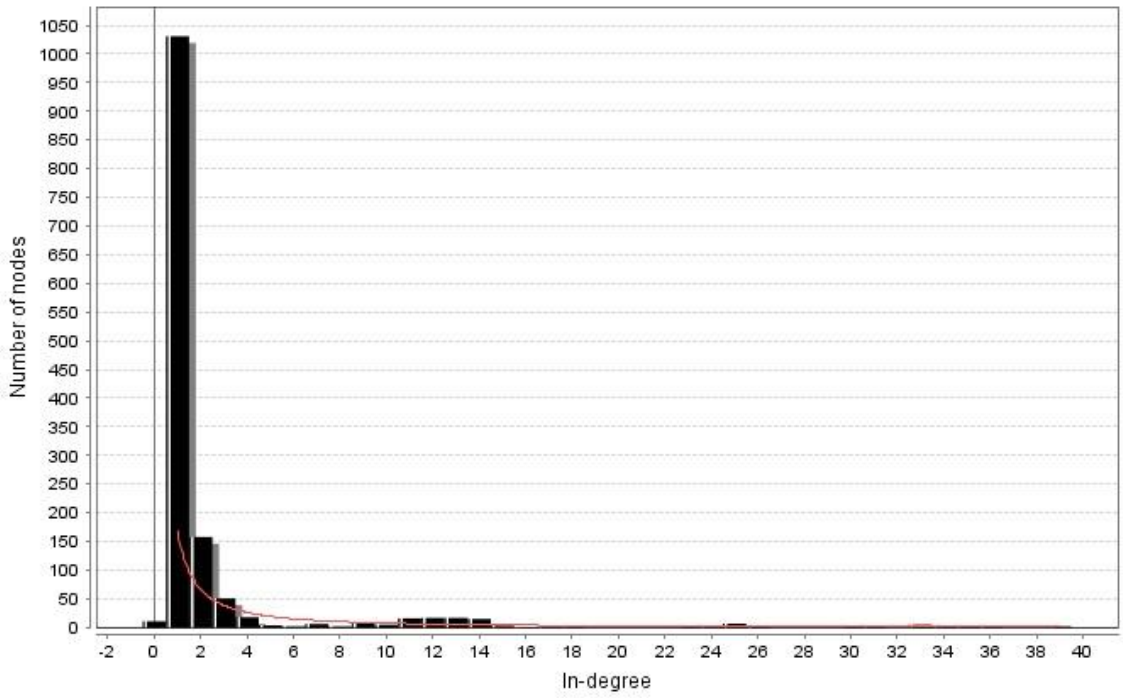
Network Diameter a graph's diameter is the largest number of vertices which must be traversed in order to travel from one vertex to another when paths which backtrack

Average number of neighbors is in general, how many neighbors does a node have.

Network Heterogeneity reflects the tendency of a **network** to contain hub nodes. In general, a hub is a node with a large degree, meaning it has connections with many other nodes.

STANFORD ANALYSIS





The data for Power Law was Calculated as follows in Cytoscape.

$y = ax^b$

a	167.83
b	**2.341**
Corelation	0.95 4
R-Squared	0.614

Average Degree	**2.6**
Average Clustering Coefficient	**0.143**
Average Path Length	**2.667**
Network Diameter	**10**
Avg. no. of Neighbors	**4.645**

Average Degree in this context implies that there are about 2.6 edges for one node in the Stanford network, or more textually that implies ***that each webpage on Stanford site would link to about 3 web pages.***

Clustering Coefficient is the ability of nodes in a graph to cluster together with neighbors.

Average Path Length is the average number of steps along the shortest paths for all possible pair of nodes in the network.

Network Diameter a graph's diameter is the largest number of vertices which must be traversed in order to travel from one vertex to another when paths which backtrack

Average number of neighbors is in general, how many neighbors does a node have.

An important thing to note here is the fact that *both the curves follow the Power Law.* The thing to ponder here is, why 2 datasets that are so unrelated, in a way, poles apart, follow Power Law?

The answer lies in the fact that both the datasets consists of nodes with many interconnections, thus *nodes that have larger number of interconnections are fewer in comparison to the nodes that have less number of interconnections*.

This is the key point of the pareto principle, clearly, rich get richer. Thus for both the datasets we have the Power Law being obeyed.

- For Slashdot, there are few users who are very popular, and thus have high number of friends and foes. Since they would have more mutual friends/foes they are more likely to have more friends/foes in future. Thus establishing the "*rich get richer*" fact.
- For Stanford Webpages, some pages are likely to be of enormous importance, such as the university website's homepage. It is but natural, that such a page will have more links to other pages and other pages would advertise this page a lot. So, establishing the *"rich get richer"* fact.

So, it is quite evident that Power Law will be followed for both the Networks.

Bibliography

- **https://snap.stanford.edu/data/** for providing datasets for Slashdot and Stanford networks that I have used for Network Analysis.
- **GEPHI** tool. The graph images would have been impossible without it.
- **Cytoscape** tool for network Analysis. All the values related to the Network were calculated using the Cytoscape tool.
- **https://en.wikipedia.org/wiki/Random_graph**
- https://en.wikipedia.org/wiki/**Barabási–Albert** model
- https://en.wikipedia.org/wiki/**Scale-free_network**
- **http://webwhompers.com/graph-theory.html**