

PLAYING WITH GENES

Through this report, I want to portray my skills in Statistical understanding, Data Analysis, Data Visualization, R, Python, Machine Learning, data wrangling, and most interestingly, story telling.

As you follow further, you'll see that I have done analysed the dataset as a **Graph** (as mathematicians call it) or as a **Network** (as data scientists call it). And, this is the most crucial aspect of my story.

Why Networks?

Now, you'd ask me, why is she emphasising this much on the "Networks" in a Data Analysis report?

"Networks are at the heart of Complex networks."

We are surrounded by systems that are hopelessly complicated. Consider for example the society that requires cooperation between billions of individuals, or communications infrastructures that integrate billions of cell phones with computers and satellites. Our ability to reason and comprehend our world requires the coherent activity of billions of neurons in our brain. Our biological existence is rooted in seamless interactions between thousands of genes and metabolites within our cells.

These systems are collectively called *complex systems*, capturing the fact that it is difficult to derive their collective behavior from a knowledge of the system's components.

We can model and analyse possibly any kind of data or information in terms of a network containing entities which are somehow linked with each other. For example:

- The network encoding the interactions between genes, proteins, and metabolites integrates these components into live cells. The very existence of this *cellular networks* a prerequisite of life.
- The wiring diagram capturing the connections between neurons, called the *neural network*, holds the key to our understanding of how the brain functions and to our consciousness.
- The sum of all professional, friendship, and family ties, often called the *social network*, is the fabric of the society and determines the spread of knowledge, behavior and resources.
- *Communication networks*, describing which communication devices interact with each other, through wired internet connections or wireless links, are at the heart of the modern communication system.

We will never understand complex systems unless we develop a deep understanding of the networks behind them.

Networks of all kinds drive the modern world. You can build a network from nearly any kind of data set, which is probably why network structures characterize some aspects of most phenomenon. And yet, many people can't see the networks underlying different systems.

The most successful companies of the 21st century, from Google to Facebook, Twitter, LinkedIn, Cisco, Apple and Akamai, base their technology and business model on networks. Networks have gained particular popularity with the emergence of Facebook, the company with the ambition to map out the social network of the whole planet.

Hence, I provide a "Network Science" approach to the Data Analysis through this report.

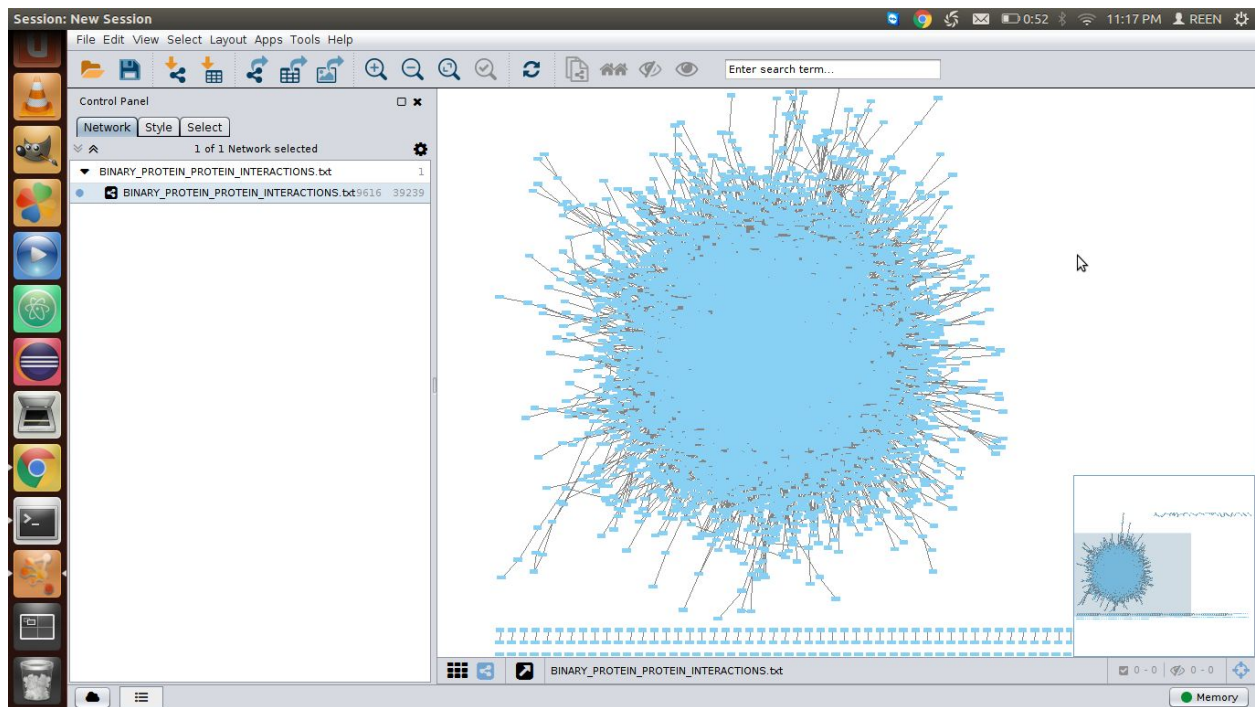
I did the analysis on the network formed using the Genes of coronary artery disease. Database source: CAD Gene Disease database.

We are given the two sets of datasets: CAD database and Human Protein Reference Database (HPRD). In CAD, we have a set of segregated genes which aren't connected and don't form any network as of now. HPRD is the protein database of all the proteins existing.

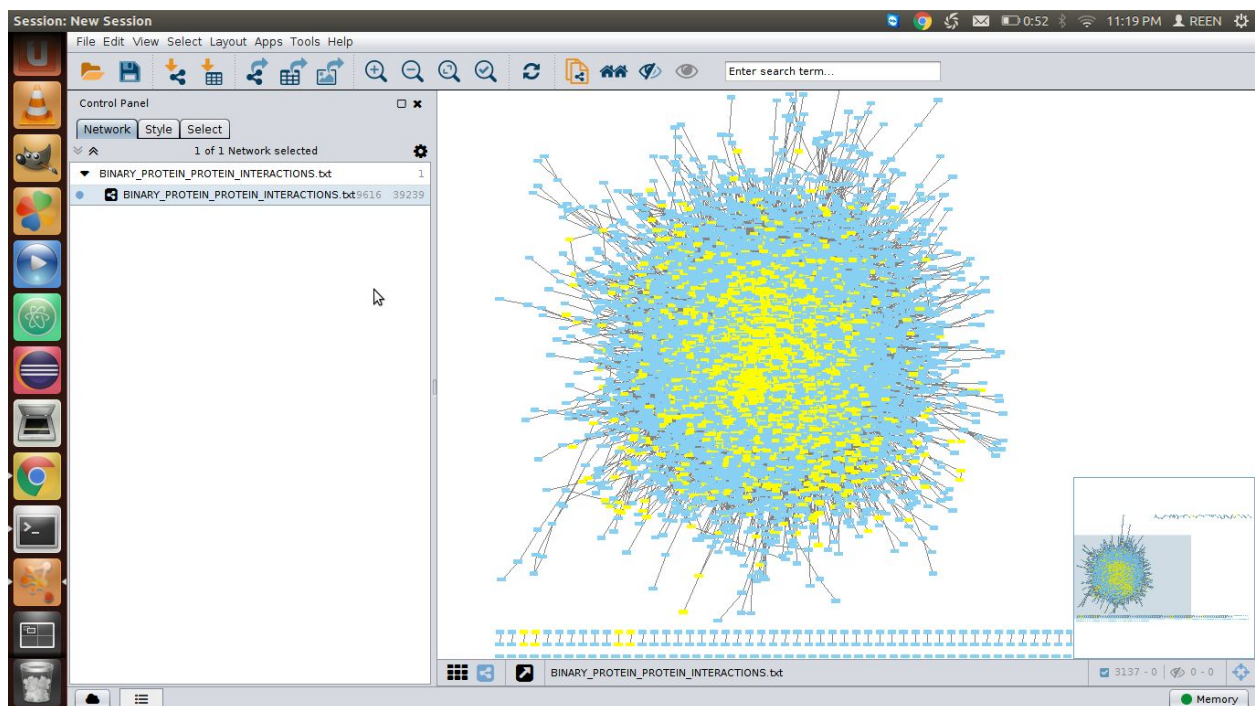
We construct a network of the dispersed CAD genes (input) by mapping it onto the Human Binary Interaction network from HPRD. In the core what we are doing is: We are looking for the CAD genes in the HPRD database and then using all the connections with those nodes in HPRD to form a network within CAD genes (as HPRD is the largest and trusted network of proteins and genes).

This is done using **Cytoscape**.

Original network:



Mapped network:



Topological Analysis:

- Degree Distribution: The *degree distribution*, p_k , provides the probability that a randomly selected node in the network has degree k . Since p_k is a probability, it must be normalized.

The degree distribution has assumed a central role in network analysis. One reason is that the calculation of most network properties requires us to know p_k . For example, the average degree of a network can be written as:

$$\langle k \rangle = \sum_{k=0}^{\infty} k p_k$$

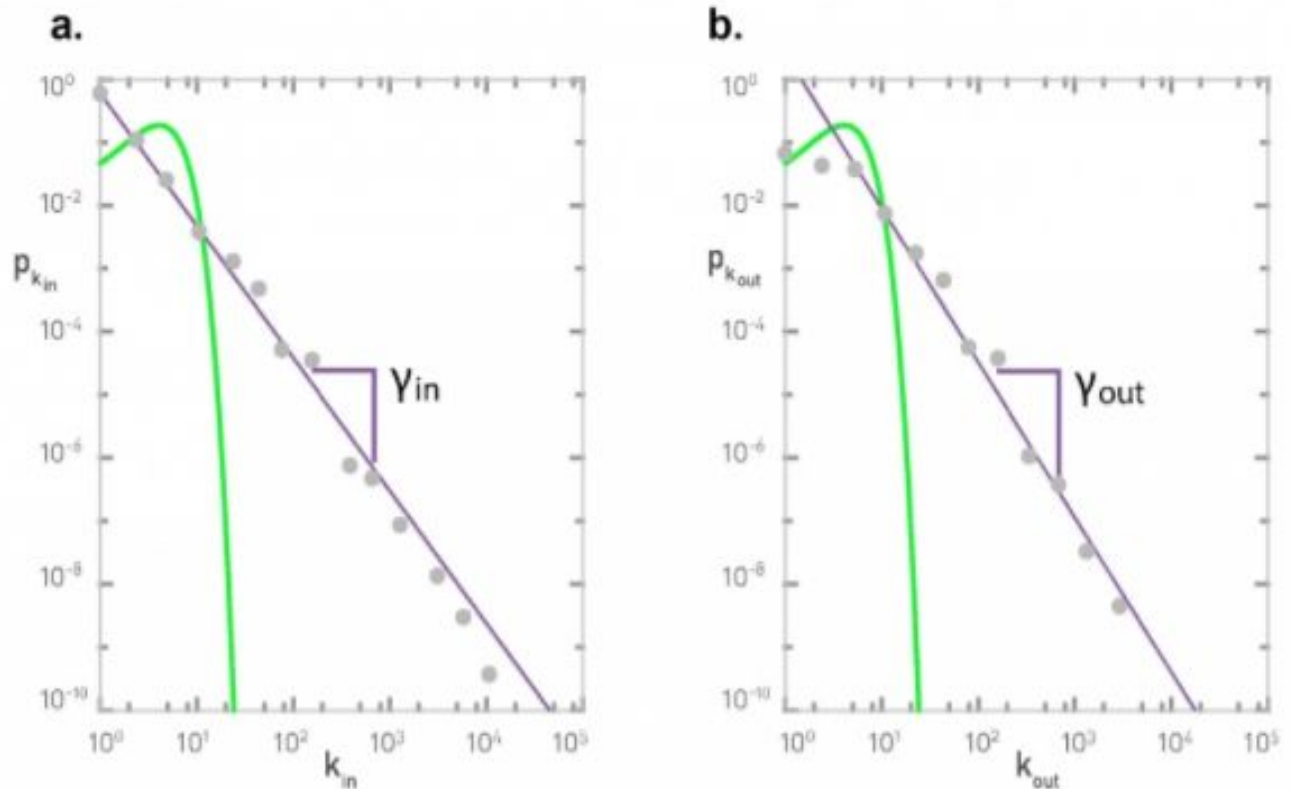
The other reason is that the precise functional form of p_k determines many network phenomena, from network robustness to the spread of viruses.

- Scale free analysis of network:

We distinguish two degree distributions: the probability that a randomly chosen document points to k_{out} web documents, or $p_{k_{out}}$, and the probability that a randomly chosen node has k_{in} web documents pointing to it, or $p_{k_{in}}$. Both $p_{k_{in}}$ and $p_{k_{out}}$ can be approximated by a power law

$$p_{k_{in}} \sim k^{-\gamma_{in}}$$
$$p_{k_{out}} \sim k^{-\gamma_{out}}$$

The incoming and Outgoing degree distribution is shown below. The degree distribution is shown on double logarithmic axis (log-log plot), in which a power law follows a straight line. The symbols correspond to the empirical data and the line corresponds to the power-law fit, with degree exponents $\gamma_{in} = 2.1$ and $\gamma_{out} = 2.45$. We also show as a green line the degree distribution predicted by a Poisson function with the average degree $\langle k_{in} \rangle = \langle k_{out} \rangle = 4.60$



A scale-free network is a network whose degree distribution follows a power law.

Significance of Power Law:

<The 80/20 Rule and the Top One Percent>

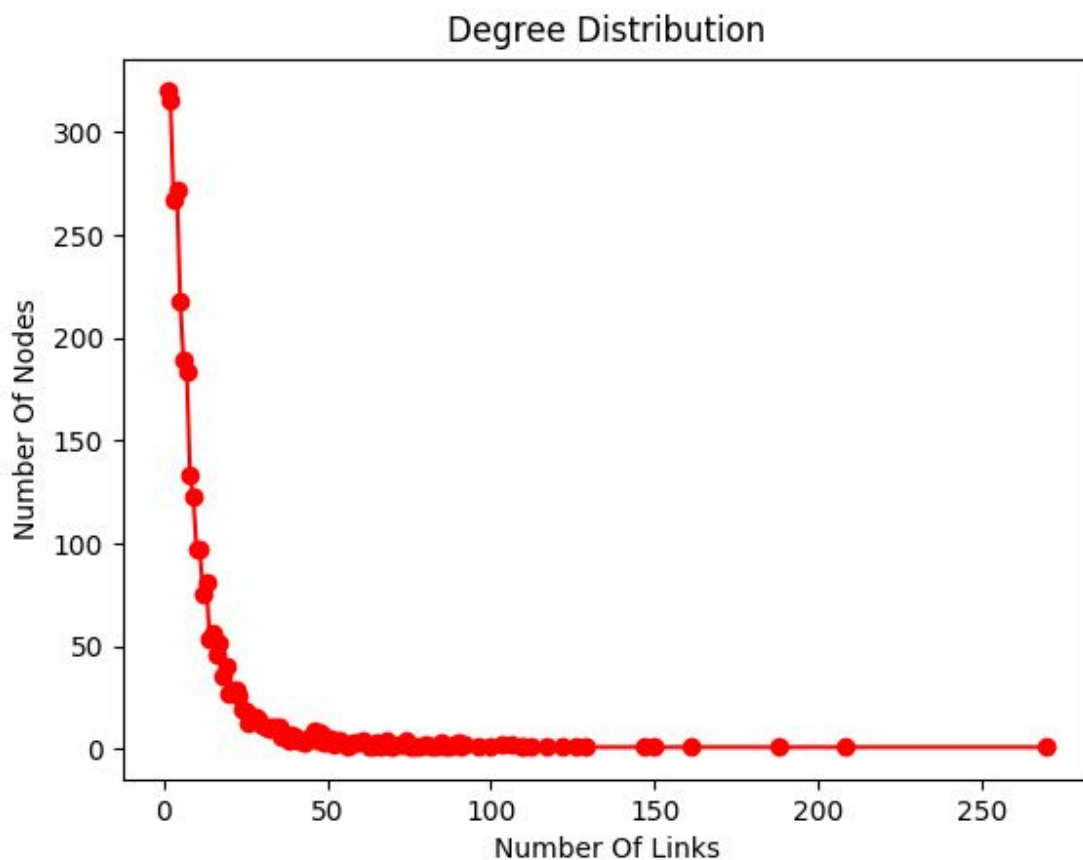
Vilfredo Pareto, a 19th century economist, noticed that in Italy a few wealthy individuals earned most of the money, while the majority of the population earned rather small amounts. He connected this disparity to the observation that incomes follow a power law, representing the first known report of a power-law distribution. His finding entered the popular literature as the *80/20 rule*: Roughly 80 percent of money is earned by only 20 percent of the population.

The 80/20 rule is present in networks as well: 80 percent of links on the Web point to only 15 percent of webpages; 80 percent of citations go to only 38 percent of scientists; 80 percent of links in Hollywood are connected to 30 percent of actors. Most quantities following a power law distribution obey the 80/20 rule.

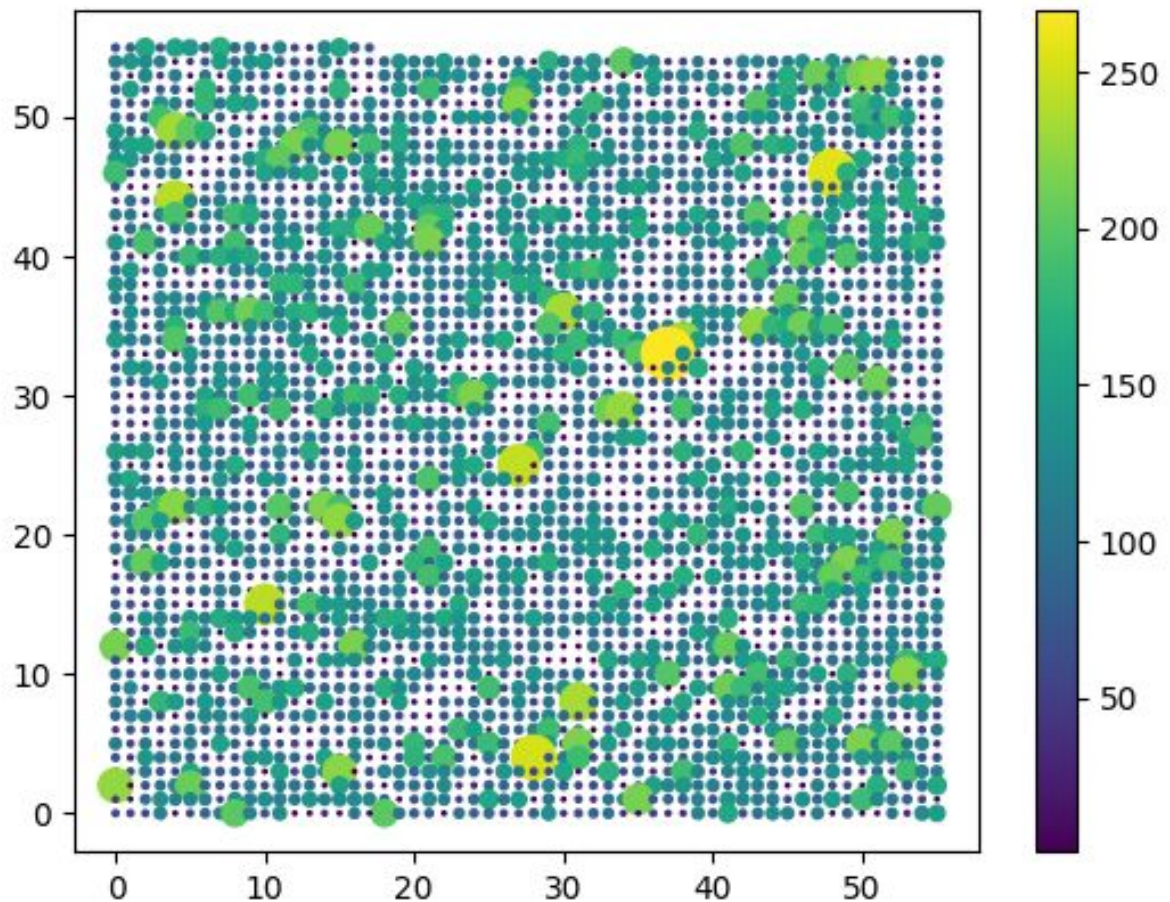
In the US 1% of the population earns a disproportionate 15% of the total US income. This 1% phenomena, a signature of a profound income disparity, is again a consequence of the power-law nature of the income distribution.

Since Power law has got such a high significance in real world network analysis, we will carry a scale-free network analysis on our database too.

We found the following degree distribution which follows the power law:



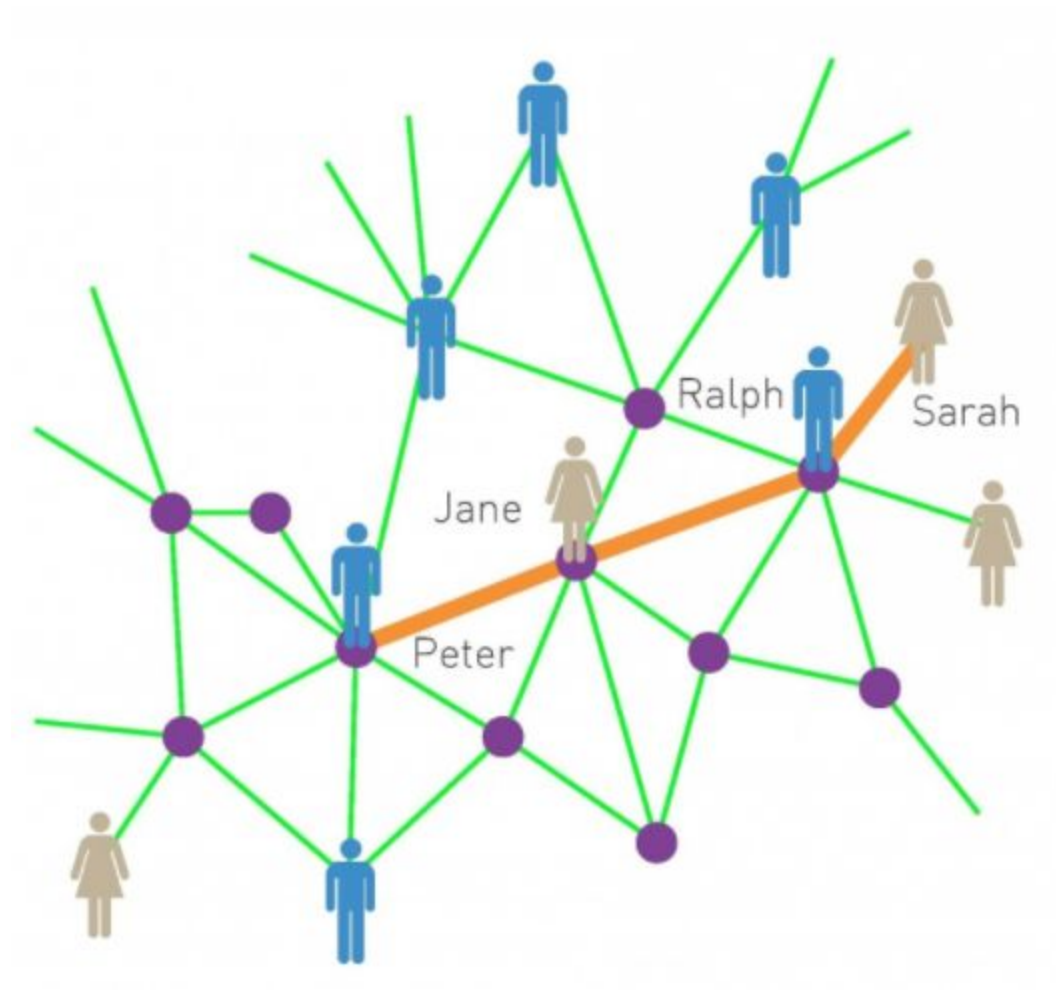
The degree of node is represented by various colours and the size of the circle represents the size of the nodes. We can conclude that there are very few nodes of high degree by looking at the figure:



Small world analysis:

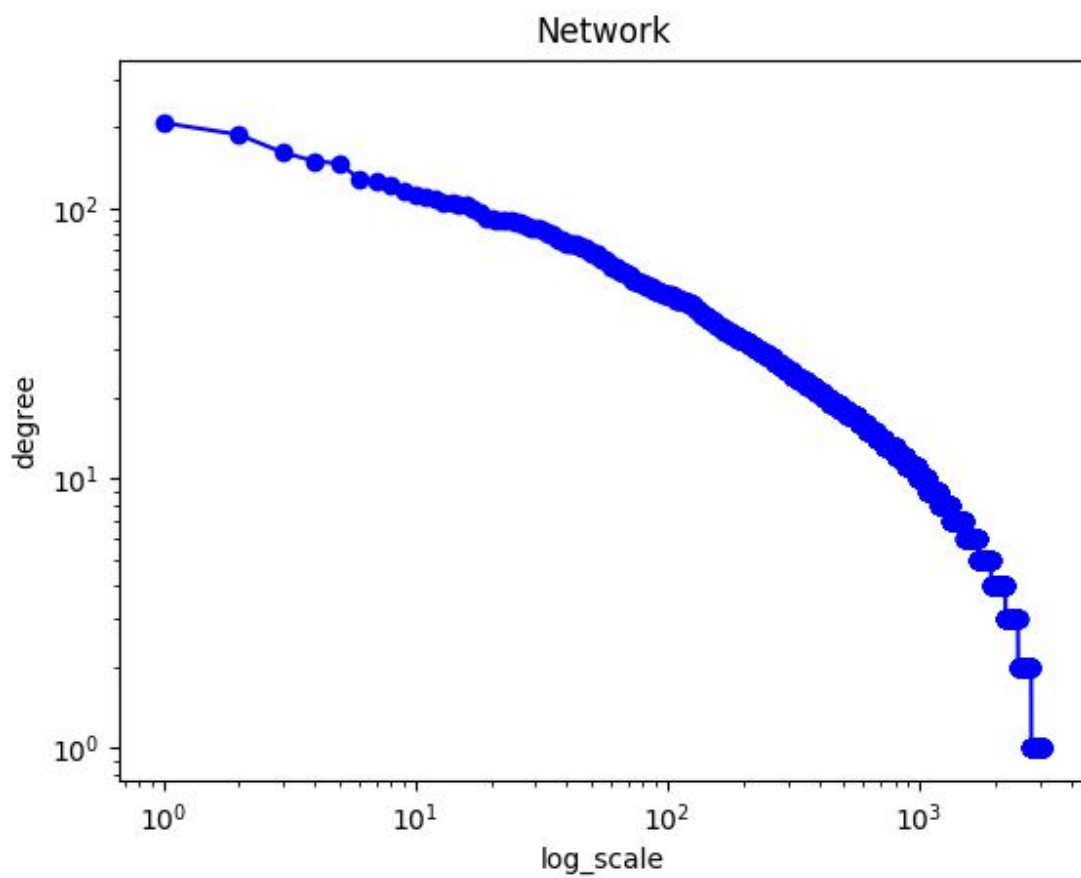
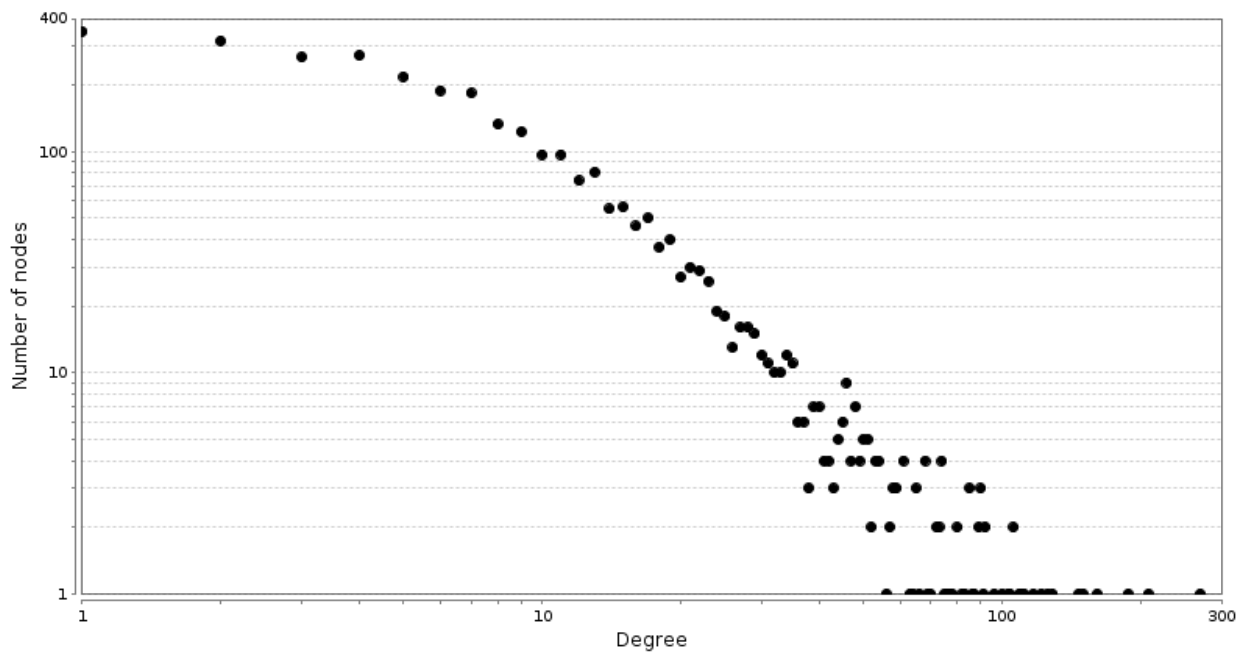
The *small world phenomenon*, also known as *six degrees of separation*, has long fascinated the general public. It states that if you choose any two individuals anywhere on Earth, you will find a path of at most six acquaintances between them. The fact that individuals who live in the same city are only a few handshakes from each other is by no means surprising. The small world concept states, however, that even individuals who are on the opposite side of the globe can be connected to us via a few acquaintances.

In the language of network science the small world phenomenon implies that *the distance between two randomly chosen nodes in a network is short*.



We carried the small world analysis on our network and observed that:

- **Clustering coefficient** came out to be **0.64**.
 - Clustering coefficient provides the information about the relationship between a node's neighbors.
 - Local clustering coefficient C_i measures the density of links in node i 's immediate neighborhood: $C_i = 0$ means that there are no links between i 's neighbors; $C_i = 1$ implies that each of the i 's neighbors link to each other.
- **Characteristic path length** came out to be **3.590**



For comparison, I am implementing the Wattz-Strogatz model on the graph having same configuration.

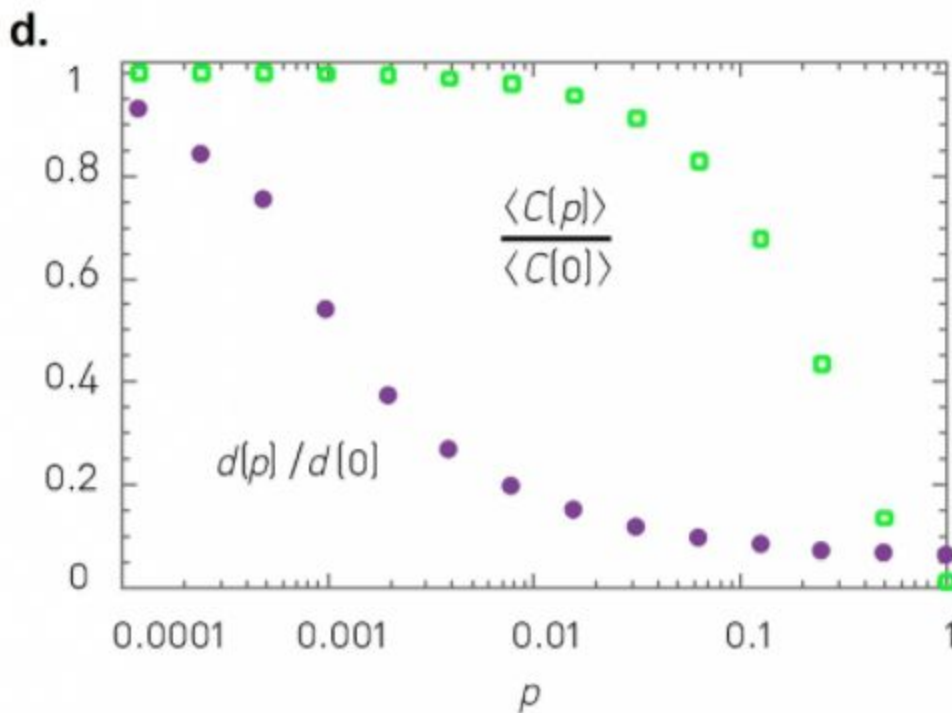
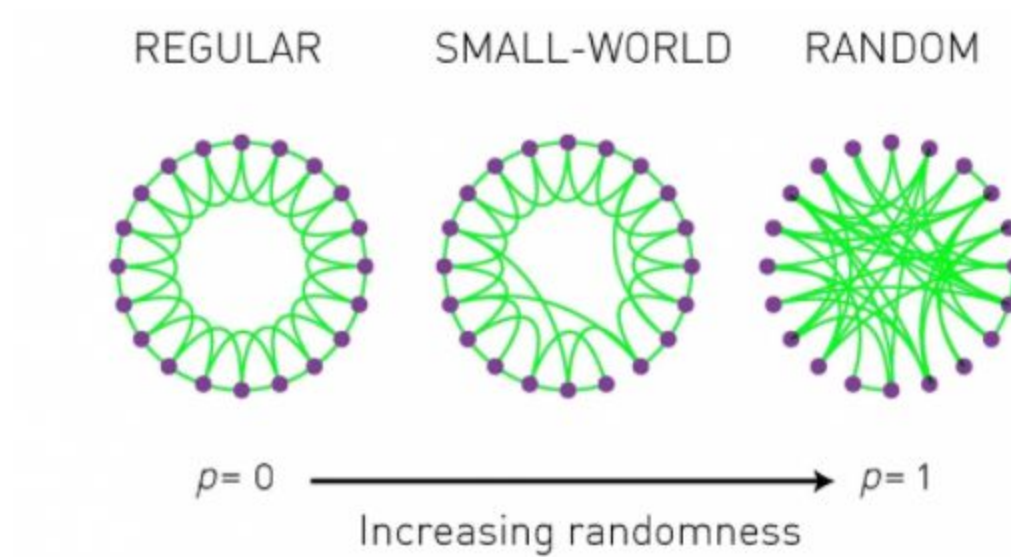
Wattz-Strogatz model:

1. Small World Property: In real networks the average distance between two nodes depends logarithmically on N , rather than following a polynomial expected for regular lattices.
2. High Clustering: The average clustering coefficient of real networks is much higher than expected for a random network of similar N and L .

The *Watts-Strogatz model* (also called the *small-world model*) interpolates between a *regular lattice*, which has high clustering but lacks the small-world phenomenon, and a *random network*, which has low clustering, but displays the small-world property. Numerical simulations indicate that for a range of rewiring parameters the model's average path length is low but the clustering coefficient is high, hence reproducing the coexistence of high clustering and small-world phenomena.

Pseudocode:

1. We start from a ring of nodes, each node being connected to their immediate and next neighbors. Hence initially each node has $\langle C \rangle = 3/4$ ($p = 0$).
2. With probability p each link is rewired to a randomly chosen node. For small p the network maintains high clustering but the random long-range links can drastically decrease the distances between the nodes.
3. For $p = 1$ all links have been rewired, so the network turns into a random network.
4. The dependence of the average path length $d(p)$ and clustering coefficient $\langle C(p) \rangle$ on the rewiring parameter p . Note that $d(p)$ and $\langle C(p) \rangle$ have been normalized by $d(0)$ and $\langle C(0) \rangle$ obtained for a regular lattice (i.e. for $p=0$ in (a)). The rapid drop in $d(p)$ signals the onset of the small-world phenomenon. During this drop, $\langle C(p) \rangle$ remains high. Hence in the range $0.001 < p < 0.1$ short path lengths and high clustering coexist in the network.



Clustering Modularity:

It indicates the degree of connectivity within the various modules of the network. Communities were detected using a networkx function and plotted according to colour. A value of 0.4804 was obtained indicating medium degree connectivity among the modules of a network also indicating there are inter-module as well as intra-module connections of roughly the same amount.

Degree Correlation:

It indicates the preference for a network's nodes to attach to others that are similar in some way. For example hubs (greater degree nodes) tend to attach to greater degree nodes in case of social media networks whereas in case of biological networks, they tend to show dissortativity, that is greater degree nodes tend to attach themselves to lower degree nodes. Even in our case, CAD protein-protein interaction network, a value of -0.033 was obtained which indicated the above mentioned result.

HUBS:

A hub is a component of a network with a high-degree node. Hubs have a significantly larger number of links in comparison with other nodes in the network. The number of links (degrees) for a hub in a scale-free network is much higher than for the biggest node in a random network, keeping the size N of the network and average degree $\langle k \rangle$ constant. The existence of hubs is the biggest difference between random networks and scale-free networks. In random networks, the degree k is comparable for every node; it is therefore not possible for hubs to emerge. In scale-free networks, a few nodes (hubs) have a high degree k while the other nodes have a small number of links.

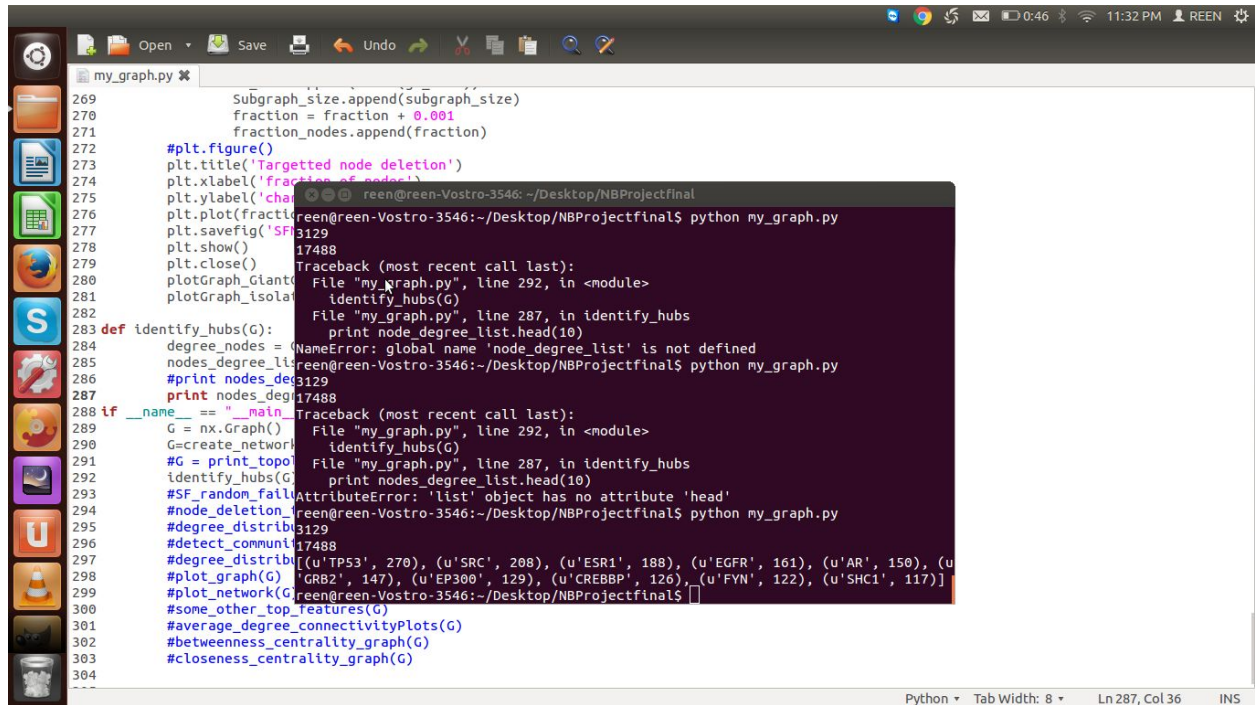
There are several crucial attributes that we can observe by studying of Hubs in scale-free networks:

- Shortening the path lengths in a network.
- Robustness and Attack tolerance:
 - During the random failure of nodes or targeted attack hubs are key components of the network. During the random failure of nodes in network hubs are responsible for exceptional robustness of network. The chance that a random failure would delete the hub is very small, because hubs coexists with a large number of small degree nodes. The removal of small degree nodes does not have a large effect on integrity of network. Even though the random removal would hit the hub, the chance of fragmentation of network is very small because the remaining hubs would hold the network together. In this case, hubs are the strength of a scale-free networks.
 - During the targeted attack on hubs, the integrity of network would fall apart relatively fast. Since small nodes are predominantly linked to hubs the targeted attack on the largest hubs would result in destruction of network in a short period of time. The financial market meltdown in 2008 is an example of such a network failure, when bankrupt of the largest players (hubs) led to a continuous breakdown of the whole system. On the other hand, it may has a positive effect when removing hubs in a terrorist network may destroy the whole terrorist group. The attack tolerance of network may be increased by connecting its peripheral nodes, however it requires to double the number of links.
- Spreading phenomenon: The hubs are also responsible for effective spreading of material on network. In an analysis of disease spreading or information flow, hubs are referred to as super-spreaders. Super-spreaders may have a positive impact,

such as effective information flow, but also devastating in a case of epidemic spreading such as H1N1 or AIDS.

Hub nodes were identified by their degree. Greater degree nodes tend to be hub nodes. The following hub nodes were obtained along with there degree. (top 10) (u'TP53', 270), (u'SRC', 208), (u'ESR1', 188), (u'EGFR', 161), (u'AR', 150), (u'GRB2', 147), (u'EP300', 129), (u'CREBBP', 126), (u'FYN', 122), (u'SHC1', 117)

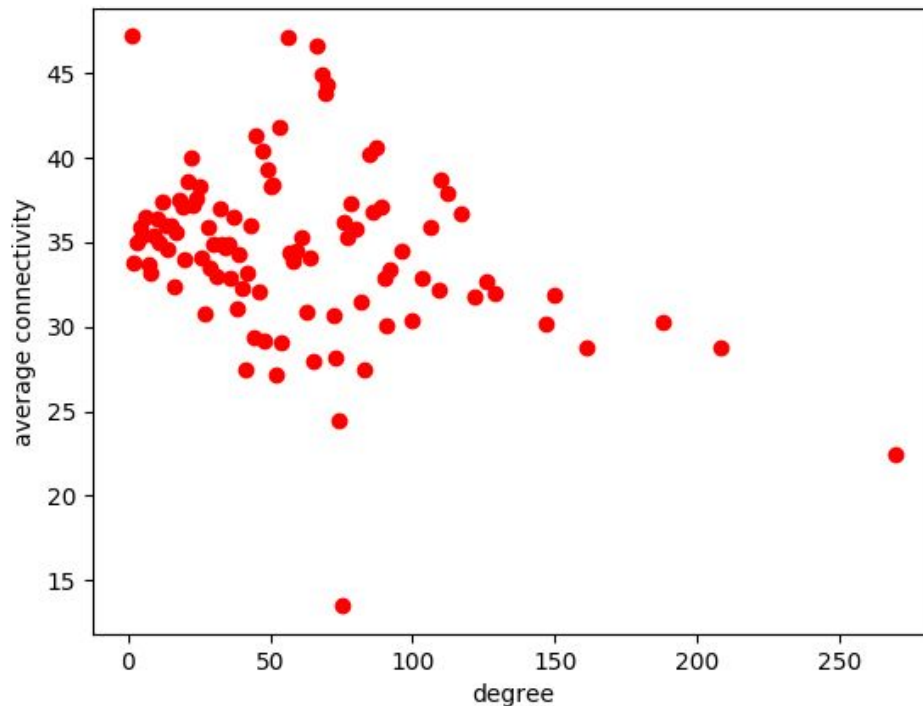
-->



```
my_graph.py
269 Subgraph_size.append(subgraph_size)
270 fraction = fraction + 0.001
271 fraction_nodes.append(fraction)
272 plt.figure()
273 plt.title('Targetted node deletion')
274 plt.xlabel('fraction')
275 plt.ylabel('chance of survival')
276 plt.plot(fraction_nodes, survival)
277 plt.savefig('SF')
278 plt.show()
279 plt.close()
280 plotGraph_Glance(G)
281 plotGraph_isolate(G)
282
283 def identify_hubs(G):
284     degree_nodes = {}
285     nodes_degree_list = []
286     #print nodes_degree_list
287     print nodes_degree_list
288
289 if __name__ == '__main__':
290     G = nx.Graph()
291     G=create_network()
292     #G = print_topological_sort(G)
293     identify_hubs(G)
294     #SF_random_failure(G)
295     #degree_distribution(G)
296     #detect_communities(G)
297     #degree_distribution(G)
298     #plot_graph(G)
299     #plot_network(G)
300     #some_other_topological_features(G)
301     #average_degree_connectivityPlots(G)
302     #betweenness_centrality_graph(G)
303     #closeness_centrality_graph(G)
304
Traceback (most recent call last):
  File "my_graph.py", line 292, in <module>
    identify_hubs(G)
  File "my_graph.py", line 287, in identify_hubs
    print nodes_degree_list.head(10)
NameError: global name 'nodes_degree_list' is not defined

reen@reen-Vostro-3546: ~/Desktop/NBProjectfinal$ python my_graph.py
Traceback (most recent call last):
  File "my_graph.py", line 292, in <module>
    identify_hubs(G)
  File "my_graph.py", line 287, in identify_hubs
    print nodes_degree_list.head(10)
AttributeError: 'list' object has no attribute 'head'
```

Average Degree Connectivity: The avg. nearest neighbour degree of nodes with degree k .



Node Deletion Studies:

Carried out to study the effects of random failure as well as targeted attack on the network. The graphs are attached in the folder.

- Characteristic path length was found to increase substantially from 3.59 to 4.3 as fraction of deleted nodes were increased from 0 - 0.04 in steps of 0.001 in case of targeted attacks whereas in case of random failures it increased only a little from 3.59 to 3.61 in the same conditions.
- The size of giant cluster also decreased, but this decrease was faster and more as compared to random failures as can be seen in the graphs.
- Size of isolated components : showed an overall decrease in targeted attacks and little less decrease in random attacks.

Another important study of data-networks is: Evaluate the value of a network, How is it related to the size of the network?

For this, I'll define and produce statistics on something we call as, "**Metcalfe's Law**".

Metcalfe's law states that the *value of a network* is proportional to the square of the number of its nodes, i.e. N^2 . Formulated around 1980 in the context of communication devices by Robert M. Metcalfe, the idea behind Metcalfe's law is that the more individuals use a network, the more valuable it becomes. Indeed, the more of your friends use email, the more valuable the service is to you.

Metcalfe's law was frequently used to offer a quantitative valuation for Internet companies. It suggested that the value of a service is proportional to the number of connections it can create, which is the square of the number of its users. In contrast the cost grows only linearly with N . Hence if the service attracts sufficient number of users, it will inevitably become profitable, as N^2 will surpass N at some large N . Metcalfe's Law therefore supported a "build it and they will come" mentality, offering credibility to growth over profits.

