# INFSCI2125: Gnutella Network Analysis

Yang Ma - xxxxx@pitt.edu



# Gnutella Structure

## Introduction

Gnutella is an unstructured peer-to-peer network. Unlike traditional Client Server networks where the central server is overloaded with query requests, peer-to-peer networks solve this issue by balancing the load among all nodes connected to the network. This feature allows the infrastructure to dynamically scale with the size of the users. Gnutella is self-organised where each node only knows the nodes around them and does not understand the topology of the network. This peer-to-peer content-sharing network made sure that the contents were distributed around the nodes and could tolerate single points of failure. However, since it's a self-organised network and users have freedom of choice on when to join and leave the network, the network structure is really dynamic and nodes will join and leave the network in an hourly manner. Moreover, Gnetella gave users the freedom of

not providing content and only joining the network as a freeloader, this further influenced the network topology.
The Gnutella network is fully decentralised so the search costs are distributed, but there is no guarantee of querying time and reliable response and the search scope is O(N). Since we focus on connectivity instead of information flow, and the network is connected with TCP connections which are bidirectional, we could assume this network is an undirected network during analysis.

## Mechanisms

*Ping-Pong Mechanism*: A node joins the network by pinging the neighbours and each neighbour will respond to a pong message upon receiving the ping message. At the same time, each neighbour will further ping its neighbours so the node would understand the local network structure.

*Query Flooding*: When querying for specific content, nodes will flood the network by sending queries to their neighbours and the neighbours will pass it on. There is a Time-to-Live(TTL) limit for the query, so it won't overload the network. The TTL limit is usually 7-hops.

*Maximum Connections*: Nodes communicate through TCP connections. Although there are no limitations, some nodes would choose to set a limit on the maximum number of connections they have. More connections mean more efficient querying so nodes would generally prefer more connections.

# Data

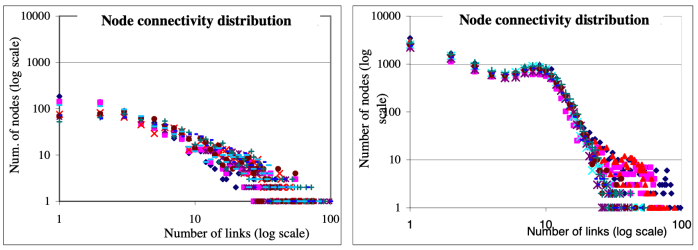
The data contains 9 snapshots of the network over August 2002. It includes the snapshots on days 4, 5, 6, 8, 9, 24, 25, 30, and 31. Since the time of the snapshot is inconsistent, the ad-hoc analysis may be unrepresentative.
The data was obtained in [1] where they used a servant as a crawler who discovers new nodes by pinging neighbours. The authors first took a serial approach with a single servant, but it turns out the crawling speed is too slow compared to the rate of network changes. So they applied multiple crawlers with a central host managing crawling tasks and found the most suitable number of crawlers to efficiently detect new nodes while avoiding overloading the network structure. However, there are still a few limitations:

1. There is no guarantee of obtaining the whole topology of the network.
2. The nodes that the crawler failed to connect to are excluded from the network constructed[1]. The failure of connection may be due to the limited number of TCP connections set of the nodes or the nodes left the network at the time the crawler trying to connect.

# Literature Review

The paper performed analysis on both topology data and network traffic data from Nov 2000 to May 2001. These data aren't published on SNAP, and we only have the topology data available from 1st August 2002 to 31st August 2002. In [1] they found that 40% of the nodes leave the network in less than 4 hours and only 25% of the nodes are alive for more than 24 hours. The early stage of the Gnutella network in Nov 2000 had the issue of traffic overhead caused by the communication

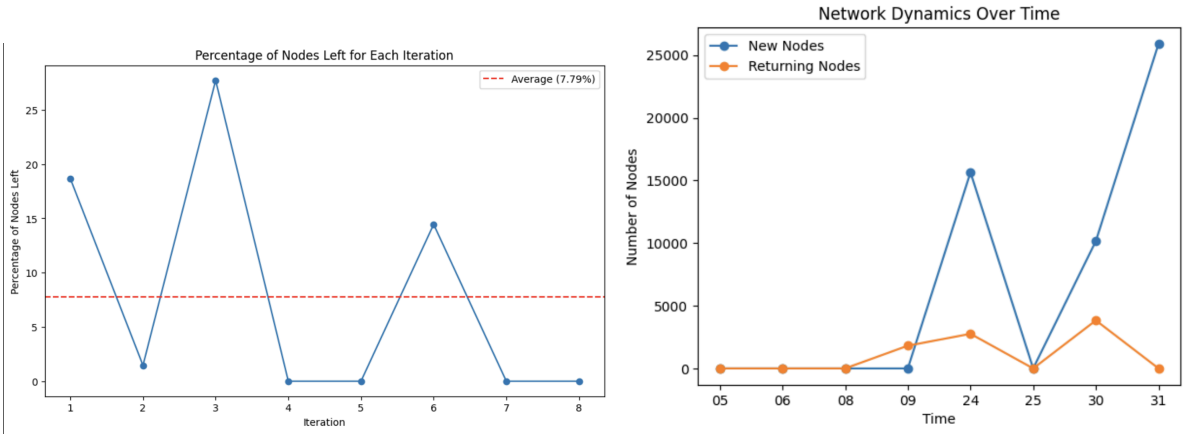
messages. This is inefficient as we want the majority of the traffic to be content-sharing traffic. A few months later after a few adjustments to the protocol, the communication traffic dropped to 8% of the whole network.




In [1], they also found that In Nov 2000, the network follows the power-law degree distribution, but a few months after, nodes there is a spike in the middle of the distribution curve, they believed its caused by the fact that people mostly prefer more connections for better query results thus the data points of low degree nodes aren't sufficient to for such power-law curve. The following plot is from [1].

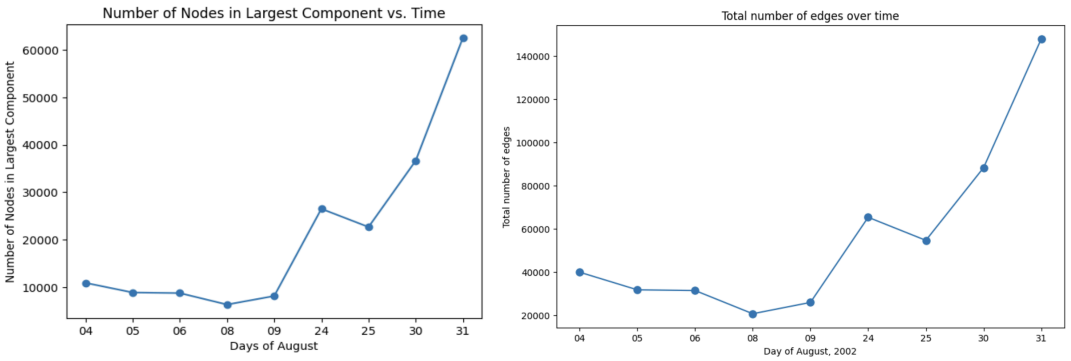
# Findings

## Churn-Rate




Based on the data in August 2002, each day around 8% of the nodes left the network on average. And that the returning nodes are relatively small compared to the new nodes who first time joining the network.

# Giant Component

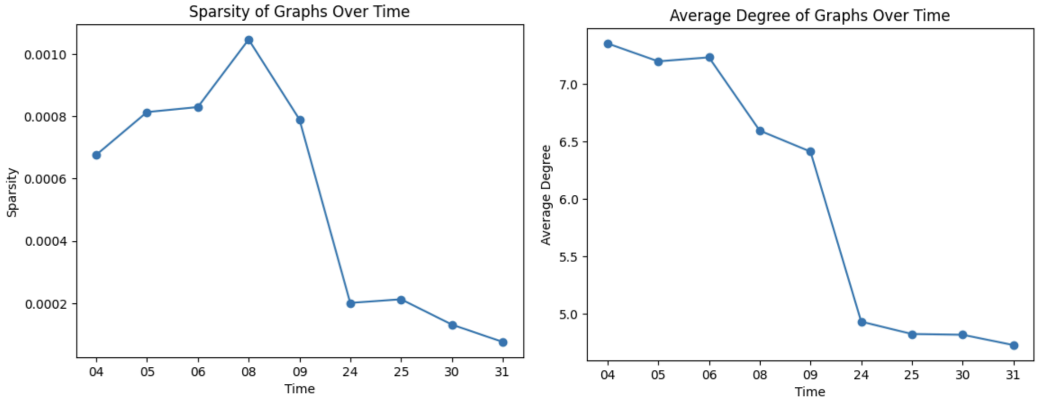



The number of nodes and the number of edges grows on a similar scale, and the network work is mostly connected.

Percentages of nodes in the largest connected component for each graph:

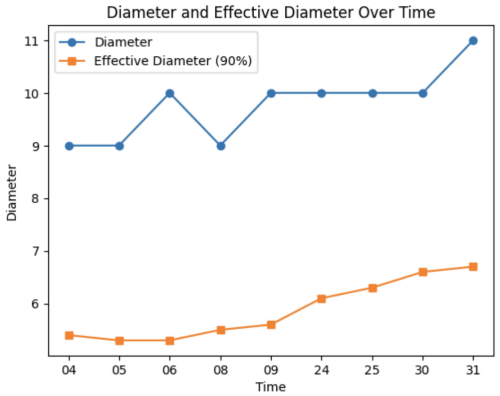
| Graph | Day 4 | Day 5 | Day 6 | Day 8 | Day 9 | Day 24 | Day 25 | Day 30 | Day 31 |
|---|---|---|---|---|---|---|---|---|---|
| Percentage | 100% | 99.95% | 100% | 99.97% | 99.88% | 99.92% | 99.89% | 99.90% | 99.96% |

# Sparsity



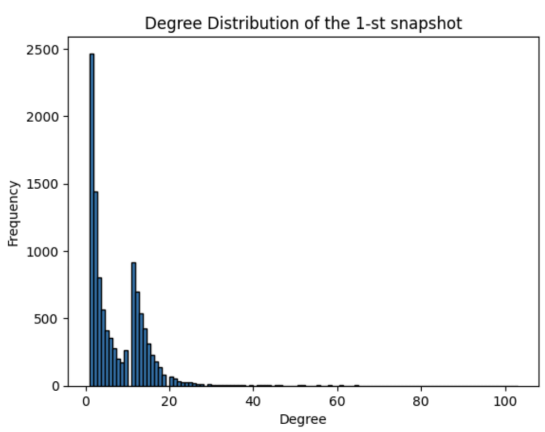

The average degree and sparsity drop as the number of nodes and edges increases. The sparsity is calculated as the number of edges over the total possible number of edges based on the number of nodes. Since the sparsity is decreasing we can tell that the network is growing denser over time. This could be due to an increase in the number of nodes and more connections being formed between existing nodes, or both.

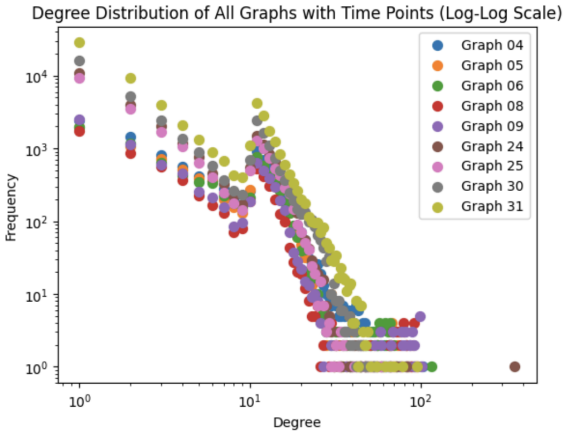
# Diameter




The gradual increase in diameter suggests that the network is expanding at a faster rate than its densification. In a P2P network, a larger diameter could increase latency, as it takes more hops to transmit data between any two nodes. However, the observed increase in diameter might not be significant enough to negatively impact the network's performance. As there is a Time-to-Live limitation of 7 hops. The diameter was above 7 the whole time and the effective diameter is still below 7. Note that this calculation is hard to reproduce due to our limited computing resources and time so we used the figures provided in SNAP.

# Power-law Distribution



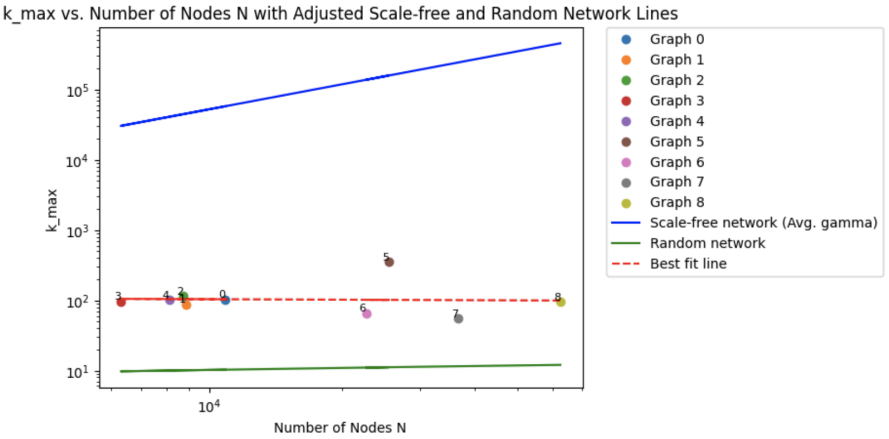

Unlike the data analysed in [1] (Data in 2000) where the network strictly follows the power-law distribution in the early stages, our data has shown a spike in degree distribution since the beginning of August 2002. Around degree 10 there is an increase in frequency. Since more connections would result in better query results, most people will try to connect to as many nodes as possible, which may lead to a spike in the degree distribution. The power-law distribution in the degree distribution indicates that the network has a scale-free property. This means that there are a few highly connected nodes (hubs) and many nodes with a smaller number of connections. In a P2P network, these hubs can improve the efficiency of resource discovery and sharing.

# Friendship Paradox

The friendship paradox is a phenomenon observed in social networks, which states that, on average, a person's friends have more friends than the person themselves. In other words, most people have fewer friends than their friends do. This paradox arises from the fact that in social networks, individuals with many friends are more likely to be friends with others, including those with fewer friends. This phenomenon has been observed in various types of networks beyond just social networks, including online social media platforms, collaboration networks, and even biological networks. It can be calculated by comparing the average number of friends that an individual has to the average number of friends that their friends have. The computed results are included in the following table, it turns out that even though the network is unstable and dynamic over time, it still persists the friendship paradox property as the average degree of neighbours is larger than the average degree of the node itself. Note that it is not as significantly larger compared to other scale-free networks.

| Day of August | 4 | 5 | 6 | 8 | 9 | 24 | 25 | 30 | 31 |
|---|---|---|---|---|---|---|---|---|---|
| Avg. degree | 7.35 | 7.20 | 7.23 | 6.59 | 6.41 | 4.93 | 4.82 | 4.82 | 4.73 |
| Avg. Neighbour degree | 14.18 | 14.38 | 14.26 | 15.78 | 15.38 | 13.12 | 12.47 | 13.05 | 13.21 |

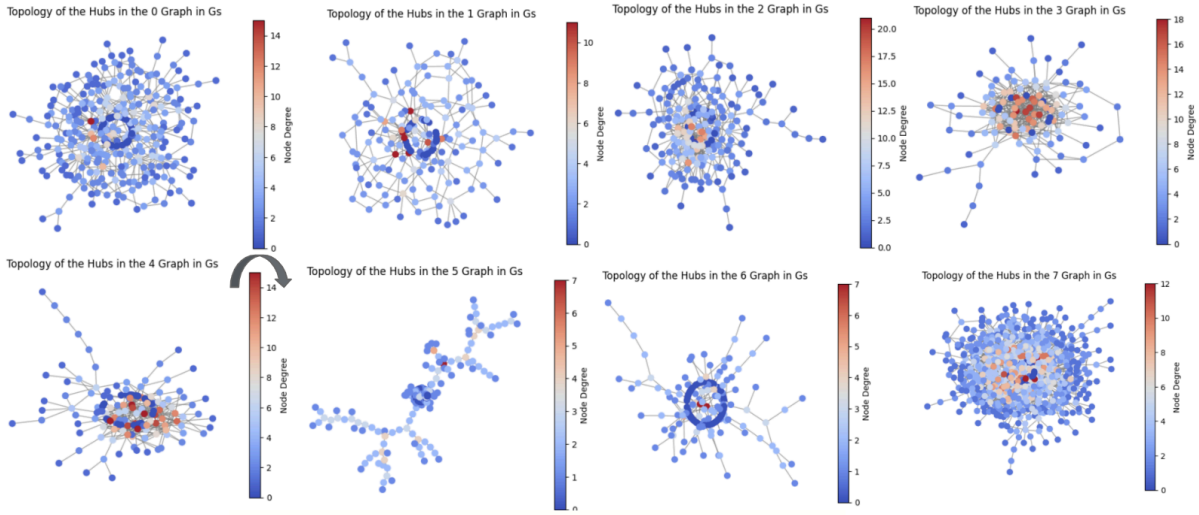
# k_max vs. N




gamma = [1.668, 1.674, 1.673, 1.754, 1.780, 1.986, 1.994, 2.036, 2.063]
Average gamma = 1.848
It shows that the node with the largest degree is consistent and does not increase as the number of nodes increases, meaning the hub is not attracting all the nodes to itself and it's a decentralized network. It proves that the network does not exhibit *scale-free* or *preferential attachment* properties, instead, it resembles a random network or a network with more evenly distributed connections among nodes.

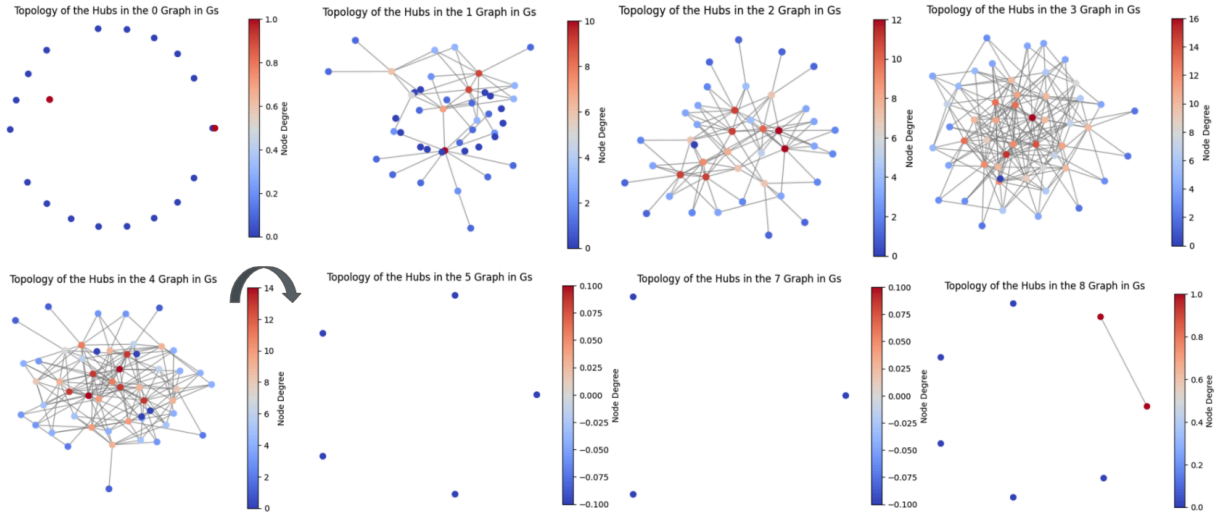
# Hubs Visualization

Here is the visualization of the hubs with threshold 20, where nodes with a degree lower than 20 are excluded from the plot for ease of visualization. More detailed plots are included in the submitted ".ipynb" file. From the visualization, we could spot that the graphs are dynamic throughout the month. The graphs in the middle of the month are not as dense as the first graph and the last graph. The networks in the middle look like it's following a disassortative property where the hubs prefer connecting to single nodes rather than hubs. Note the giant leap between graph 4 and graph 5, where 15 days of data are missing, that is why we see such a sudden change in network structure.



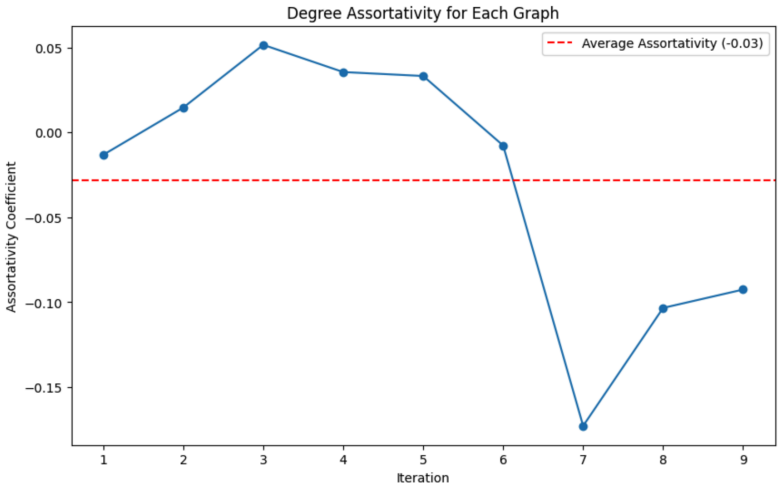

Days: ['04', '05', '06', '08', '09', '24', '25', '30', '31']

Moreover, although the network is increasing, there is no guarantee that the hubs will grow along the number of nodes. Below is the visualization of the network with a threshold of 50, nodes with degrees lower than 50 are excluded. This property is also shown in the previous section.




Days: ['04', '05', '06', '08', '09', '24', '25', '30', '31']

# Assortativity

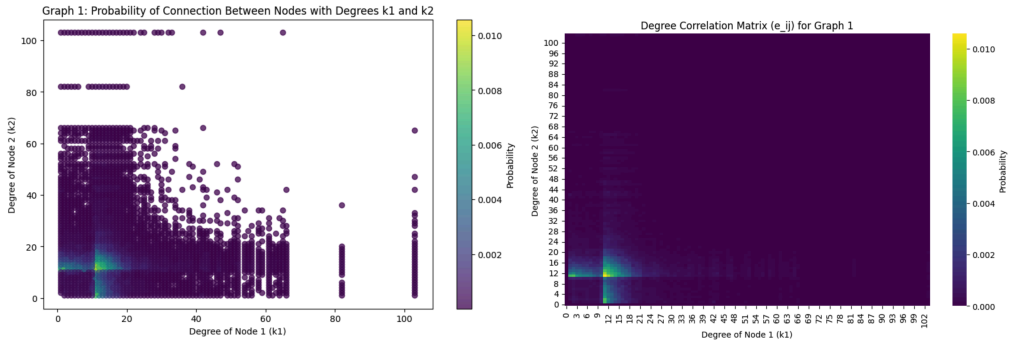



To figure out whether the hubs are more likely to connect to individual nodes or other hubs, we calculated the assortativity coefficient of the networks, it shows that the average is around -0.03 which means the network is *neutral* and the wiring is completely random in general. It's interesting to note that if we only look at graphs from 6 to 9 we could see the negative assortativity correlation, this could be seen in the visualization of the hubs with threshold 20 in the previous section. In those plots, the hubs show low connectivity among each other, with a structure that looks like branches. We could assume that it is a neutral network overall, but the network is dynamic and the assortativity slowly and slightly shifts around 0.

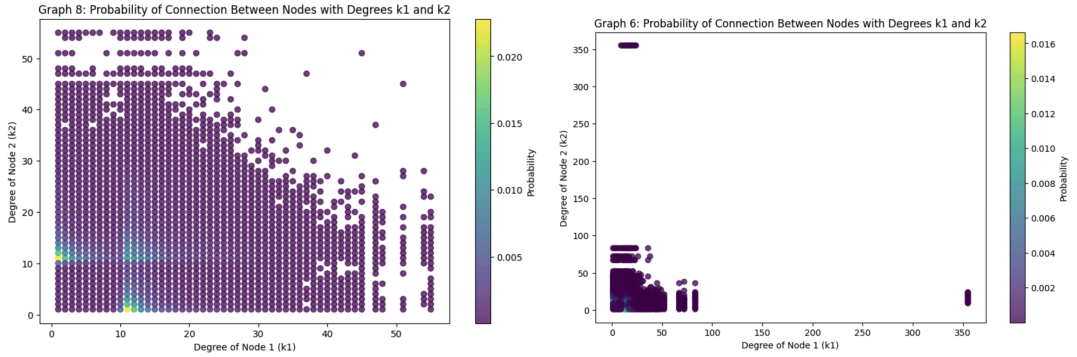
To further verify our result, we investigate the connection probabilities between nodes with specific degrees in the network. To accomplish this, the joint degree distributions of the graphs were computed, which quantifies the likelihood of two connected nodes having particular degrees. This involved iterating through all the edges in the input graph and tallying the occurrences of each pair of degrees present as endpoints of an edge. By normalizing the joint degree distribution, we derived the probabilities of connections between nodes of varying degrees.

With the joint degree probabilities computed, a scatter plot was created where the x and y coordinates represent the degrees of node pairs. The colour of each point in the scatter plot signifies the probability of a connection between nodes with the corresponding degrees. This approach was applied to each graph in the dataset, generating a series of scatter plots that offer a visual representation of the joint degree probabilities. This method enables a comprehensive comparison of connection probabilities across the entire range of graphs, revealing valuable insights into the network's structure and connectivity patterns.

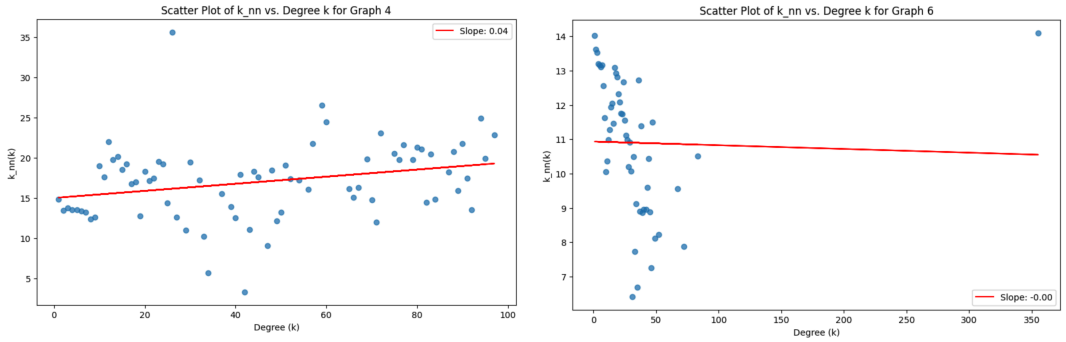
Below are the plots of the first snapshot(scatter plot and heatmap), all plots follow a similar distribution so only some of the graphs are included in the report. It is shown that the nodes with a degree around 10 are more likely to connect to other nodes. The density of links is symmetric around the average degree (6~15). The links are mostly randomly distributed, but the high-degree nodes are unlikely to connect to each other, indicating the network is close to disassortative.




The points are evenly distributed throughout the plot, especially for graph 8, which corresponds to the left-bottom network visualization in the previous section. For graph 6, which corresponds to visualization 5 in the previous section, demonstrates a more clustered distribution among low-degree nodes, close to the disassortative property.




This can be further verified with the k_nn vs. k where k_nn is the average degree of the node's neighbour. The slopes of all plots are close to 0, more graphs are included in the “ipynb” submitted under the degree assortativity section. Despite the slope of the line of the best fit(found by linear regression) for graph 6 being zero, the scatter points look more like disassortative relation.




These results explained why there is a spike in the degree distribution - most of the links lie around the nodes with degrees around 10.

# Future work

Future focus could be acquiring a longer series of data to have a more consistent analysis of the network. In the paper, the number of nodes in the largest network grow from around 2000 hosts in Nov 2000 to 50,000 hosts in May 2001. However, in our report, the plot of such data in August 2002 changes shows a similar growth rate and it also went from 1,000 to 60,000. Some more interesting findings could be found if more consistent data is acquired. Although a peer-to-peer network is robust,

it is still vulnerable to popularity. The network became unpopular and almost died out by the end of 2008. Moreover, community detection could provide some interesting results but due to the limited computation resources, it requires a much longer time to get meaningful results.

In the GitHub repo [3], the author performed a SIR simulation on the topology snapshot of 09, August 2002. It's worth noting that the graph is directed when we consider information flow, so it might be interesting to form an analysis of the directed network as well. Moreover, it may be interesting to perform the simulation on the dynamic network as the nodes leave and come back to the network if we have more consistent data over time.

# Conclusion

As the number of nodes in the network increases, the graph gets denser, but the diameter or the efficient diameter of 90% is increasing, meaning the rate that which the network is expanding is faster than the rate it densifies or a certain mechanism in the network is leading to this behaviour. This illustrates the fact that 10 per cent of the contents in the network are always unreachable unless the TTL limit is raised, and we need to balance the trade-off between the TTL and traffic overhead.

Moreover, we also found that although the degree distribution is power-law alike, it is not a scale-free network and the hubs will not expand with the number of nodes.  The friendship paradox holds in the Gnutella network, where the neighbours' degree doubles or triples the node's degree. This ratio is not as large compared to scale-free networks.

In conclusion, the high-degree nodes in the Gnutella network avoid connecting to each other, resulting in a dynamic network that looks like a combination of a neutral and a slightly disassortative network where most of the links are among nodes whose degrees are close to the average degree(ranging from 6 to 15). The network is slightly disassortative in many snapshots. Even when the assortative correlation is slightly positive, the distribution pattern persists with only a few more links connecting high-degree nodes and most of the links are densely surrounding the average degree. From the calculated assortativity, it looks like the assortativity is shifting around 0 (neutral). However, from the distribution plots, it more looks like the dynamic network could change in whichever way as long as high assortativity is avoided.

# References

[1] Ripeanu, M., Foster, I. and Iamnitchi, A., 2002. Mapping the gnutella network: Properties of large-scale peer-to-peer systems and implications for system design. *arXiv preprint cs/0209028*.
[2] Leskovec, J., Kleinberg, J. and Faloutsos, C., 2007. Graph evolution: Densification and shrinking diameters. *ACM transactions on Knowledge Discovery from Data (TKDD)*, *1*(1), pp.2-es.
[3] SIR_on_Gnutella [GitHub]
https://github.com/maufadel/SIR_on_Gnutella/blob/master/Epidemic%20Process%20on%20Gnutella%20p2p%20network.ipynb

# Appendix - Submitted Code Explanation

The submitted code is included in the file “Submission_Gnutella.ipynb” and “submission_gnutella.py”. Here is the link to the notebook via Colab: xxxxx
The code contains some draft code and useless results that I did not include in the report.