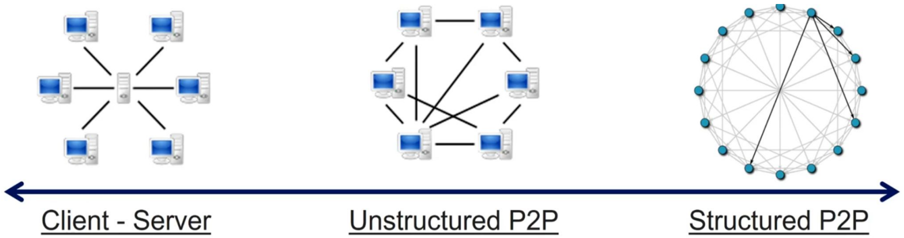
INFSCI2125 Final Project: Gnutella P2P Network Analysis

Yang

Gnutella Network – Unstructured P2P

GNU +

=



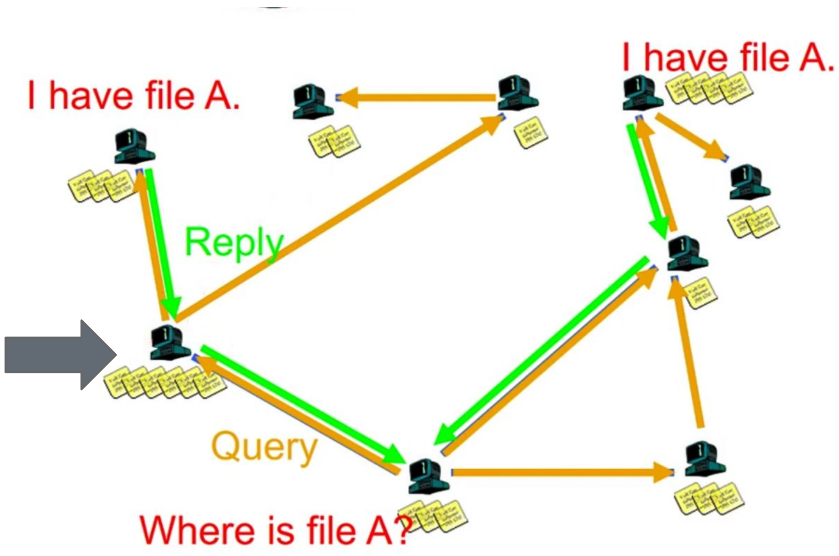
Technical goals:

- Self-organization(vs. centralized)
- Availability (vs. failure, censorship)
- Load balancing (vs. freeloading, locality)
- Anonymity (vs. monitoring)

Challenges:

- Node churn
- Control overheads
- Malicious nodes

Gnutella Network – More Details



Mechanism - Query Flooding:

- Join the network: ping-pong neighbours
- Publish: no need
- Search: flood query
- Fetch : direct download from peer

TTL-limited search (usually 7 hops)

- Query dies after
- Only works well for common objects

Pros:

- Fully decentralized
- Search cost distributed

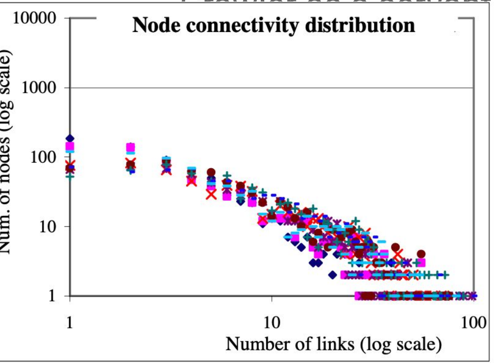
Cons:

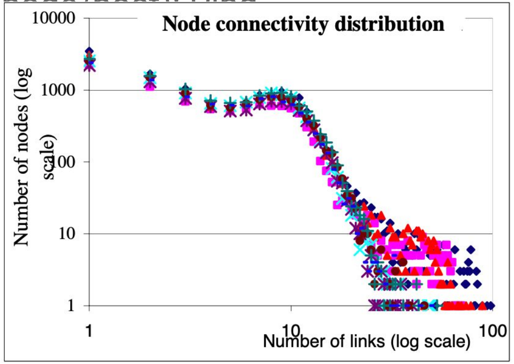
- Search Scope is O(N)
- Search time can’t be determined
- Nodes often leave, network unstable

Literature Review – Mapping the Gnutella Network

Data

Crawler as a servent(node/host): Ping



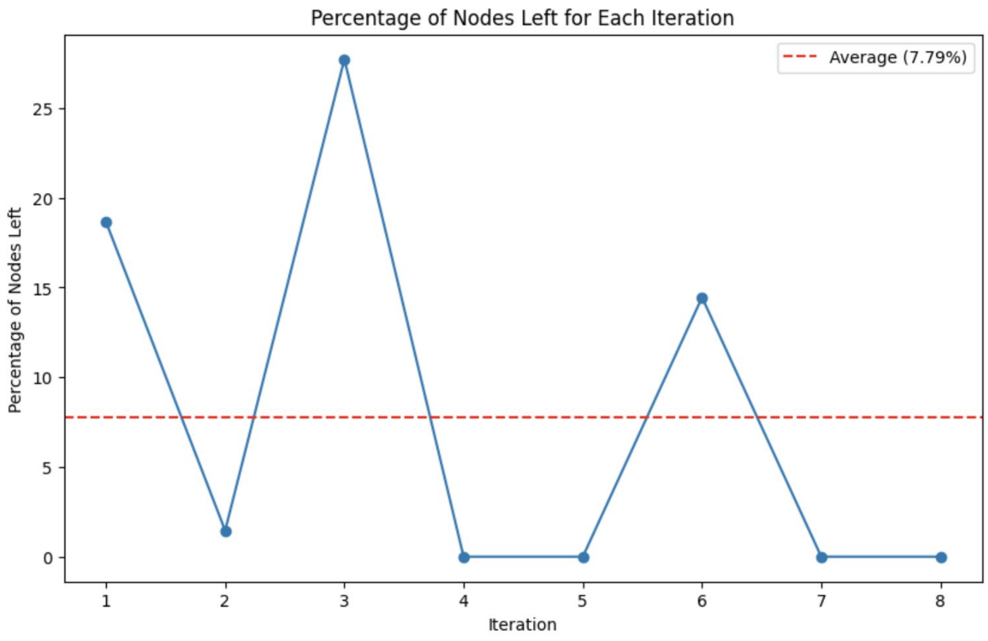


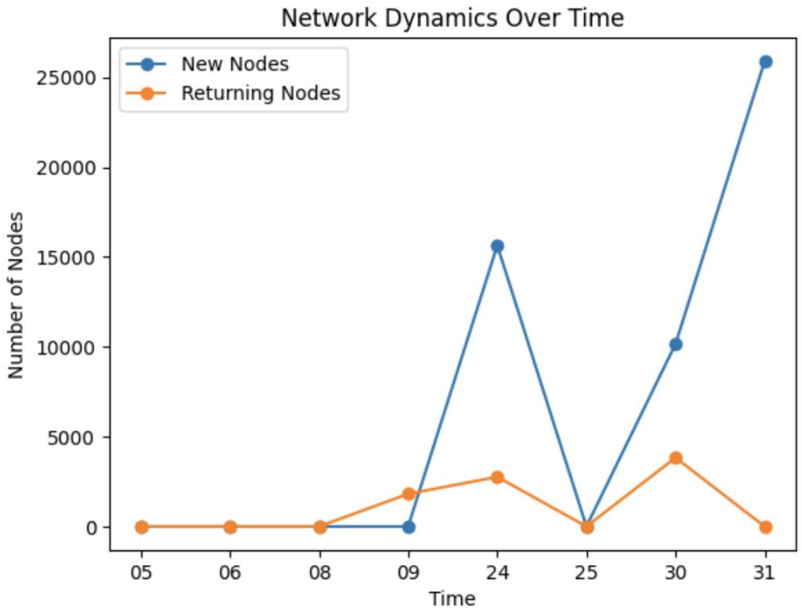
because the node has set limited number of TCP connections, or the node has left the network.

Findings:

1. Dynamic network: 40% of the nodes leave the network in less than 4 hours, only 25% of the nodes are alive for more than 24 hours.
2. Massive overhead traffic in the early stage with more than 55%(Nov 2000). Dropped to 8% later with newer implementations(June 2001).
3. Power-law distribution before Nov 2000; too few nodes with low connectivity to form a pure power-law network afterwards

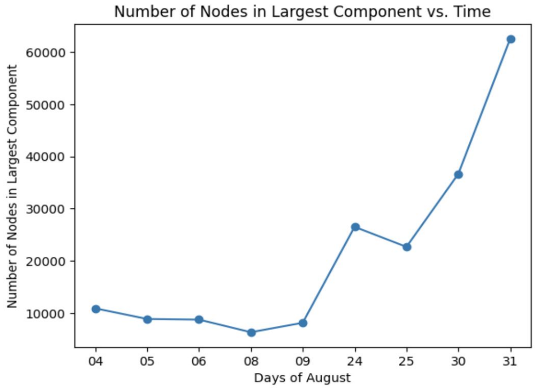
Churn Rate

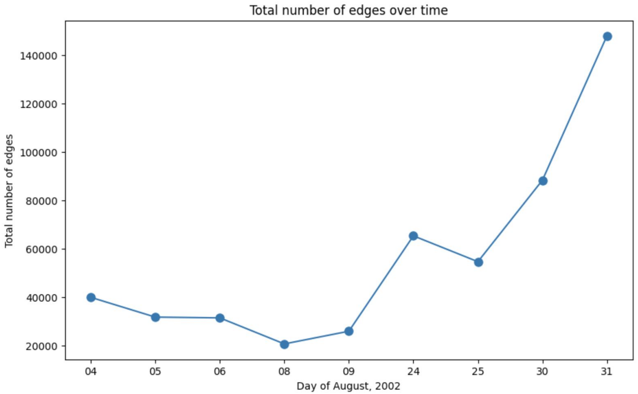




Days: `['04', '05', '06', '08', '09', '24', '25', '30', '31']`

Giant Component





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

Graph 0: 100.00%

Graph 1: 99.95%

Graph 2: 100.00%

Graph 3: 99.97%

Graph 4: 99.88%

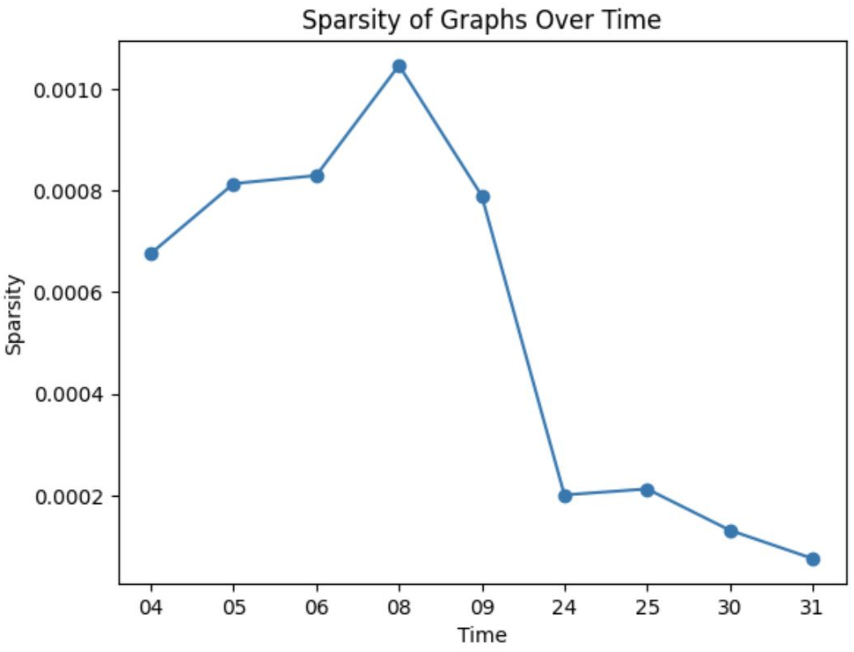
Graph 5: 99.92%

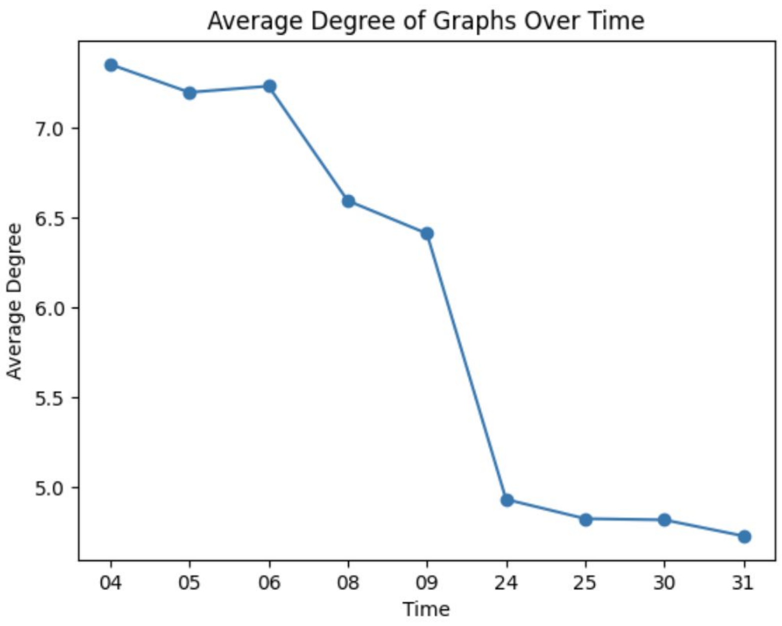
Graph 6: 99.89%

Graph 7: 99.90%

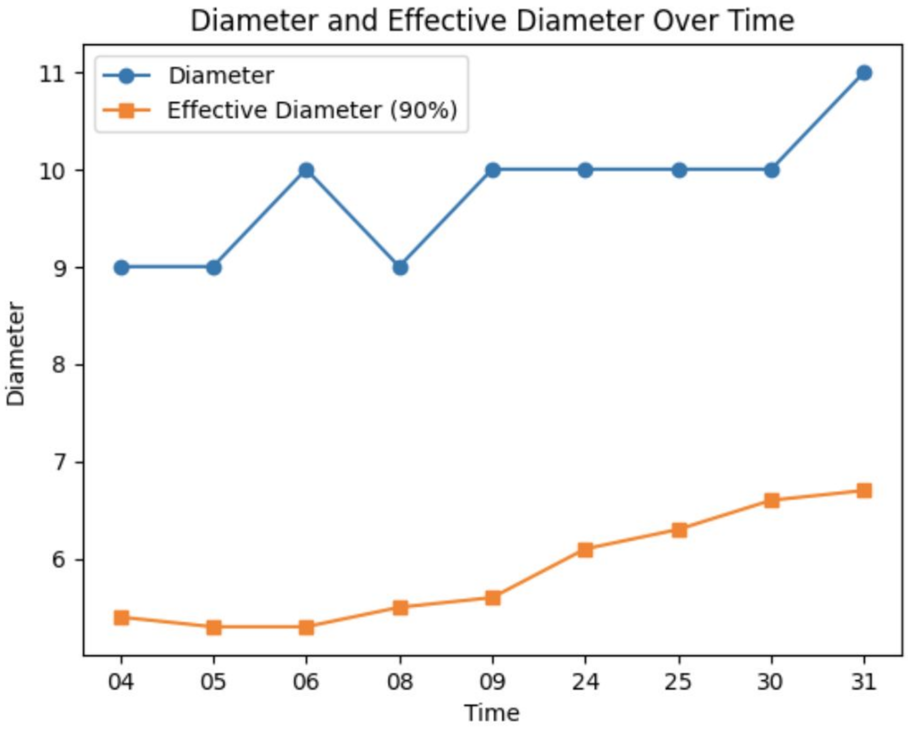
Graph 8: 99.96%

Sparsity


Sparsity of Graphs Over Time
Sparsity
0.0010
0.0008
0.0006
0.0004
0.0002
04
05
06
08
09
24
25
30
31
Time

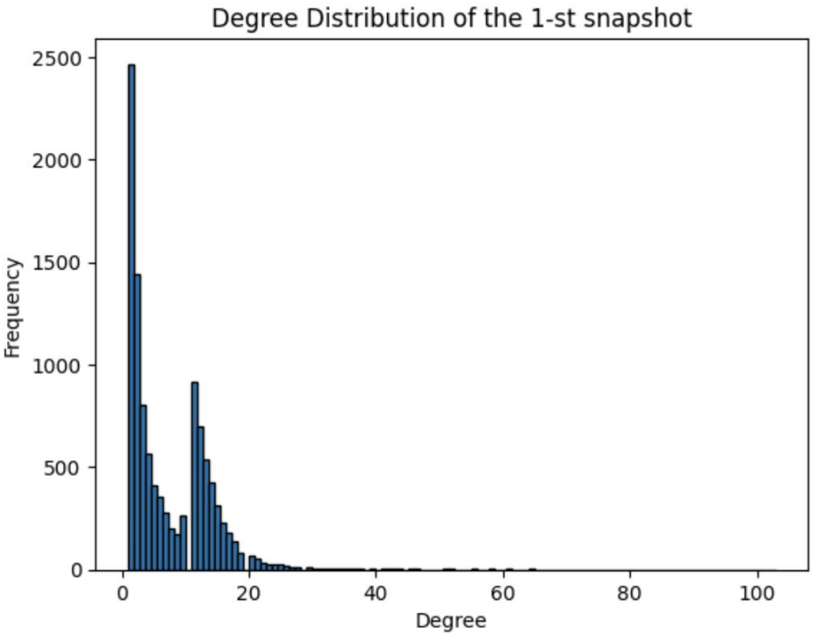

Average Degree of Graphs Over Time
Average Degree
7.0
6.5
6.0
5.5
5.0
04
05
06
08
09
24
25
30
31
Time

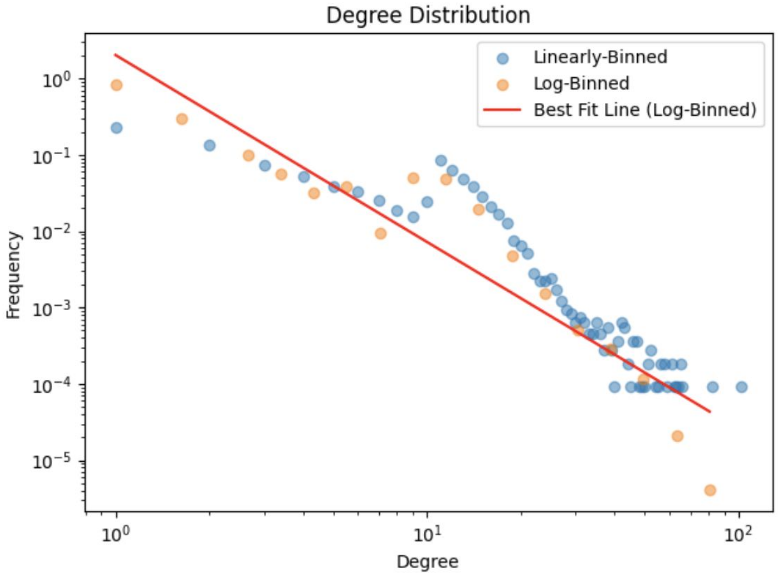
Diameter



- Network expanding faster than densification
- No effects on query due to TTL 7 hops limitation.

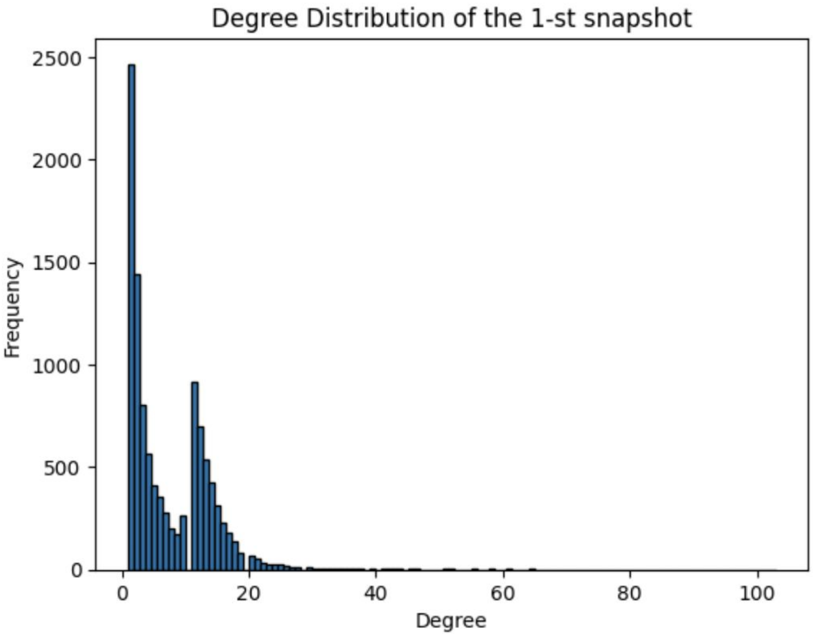
Scale-Free: Power-law Distribution

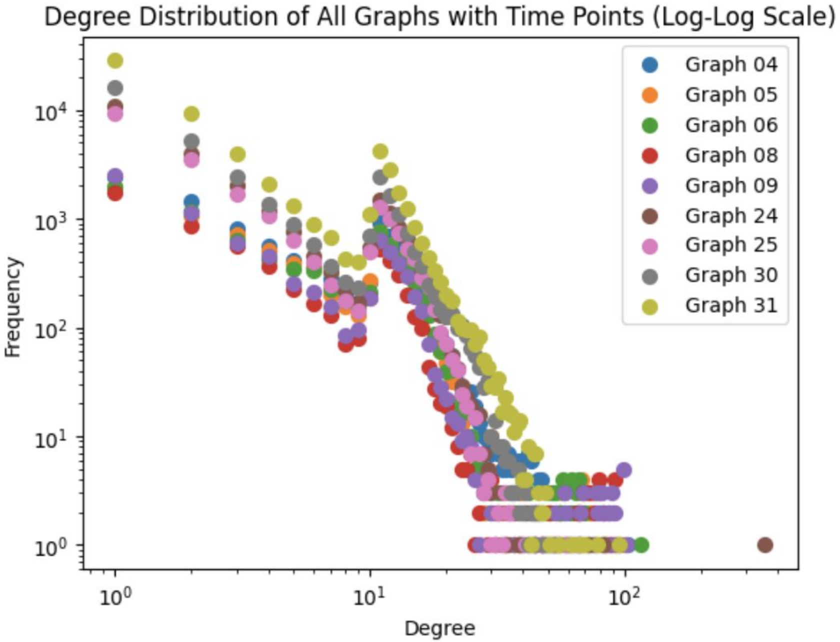




Slope: -2.449

Scale-Free: Power-law Distribution





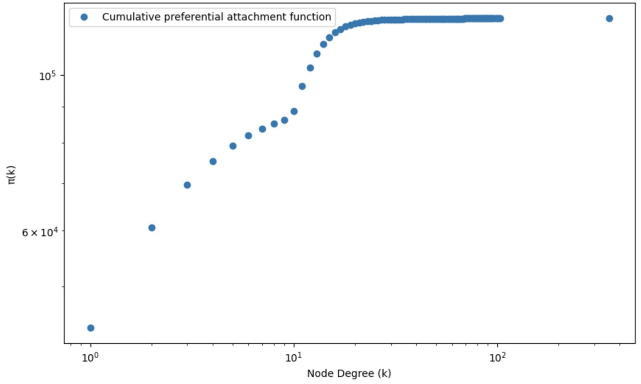
Scale-Free: Preferential Attachment

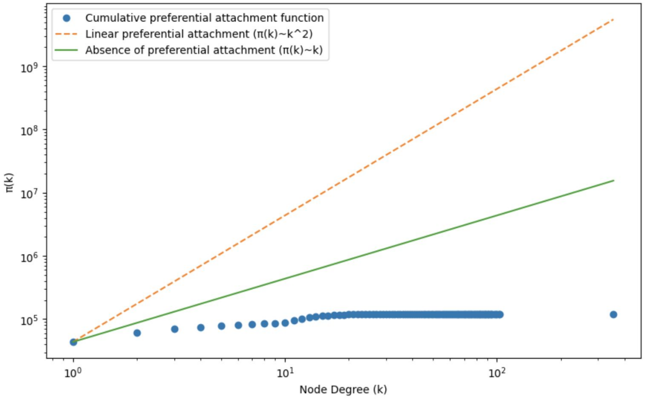
Days: `['04', '05', '06', '08', '09', '24', '25', '30', '31']`

Hypothesis 1
The likelihood to connect to a node depends on that node's degree *k*. This is in contrast with the random network model, for which *Π(k)* is independent of *k*.

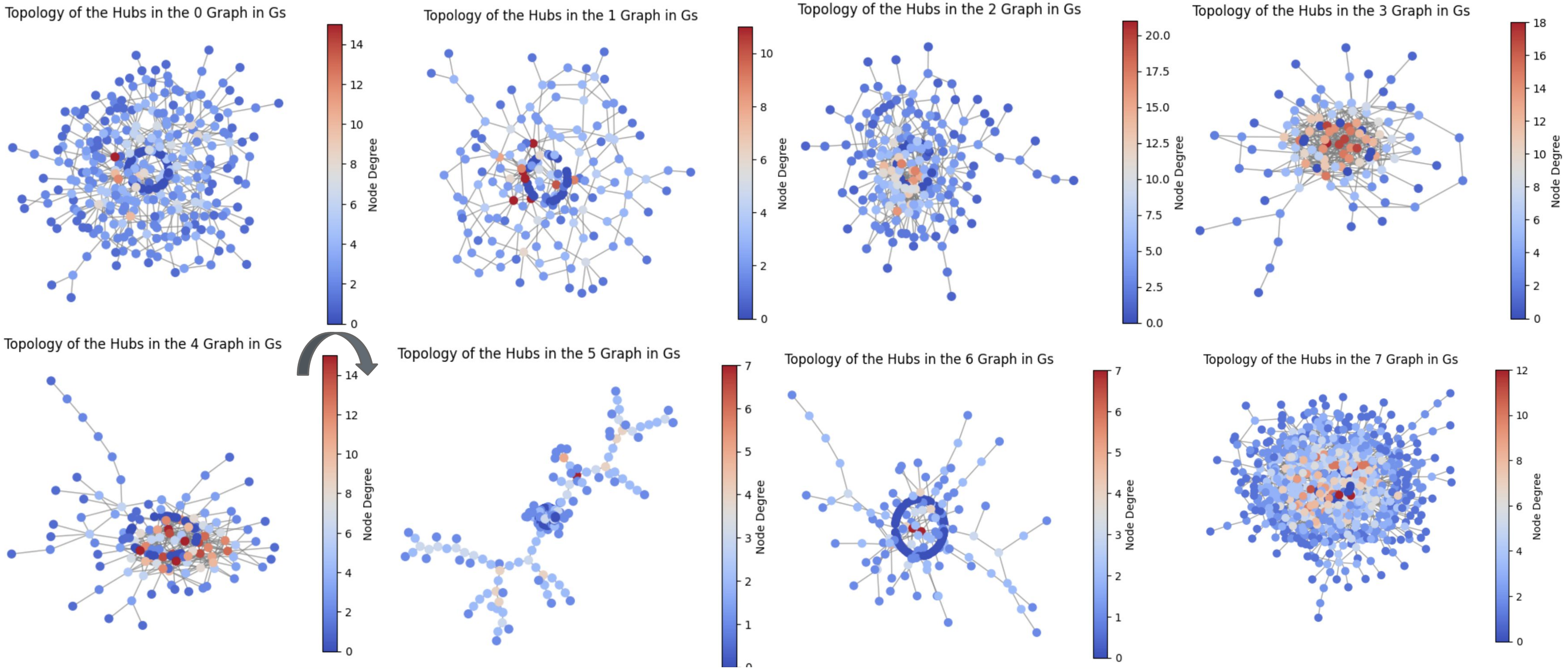
Hypothesis 2
The functional form of *Π(k)* is linear in *k*.

We detect k-dependance, thus it in line with H1.
Grows slower than (π(k)~k2), thus sublinear.



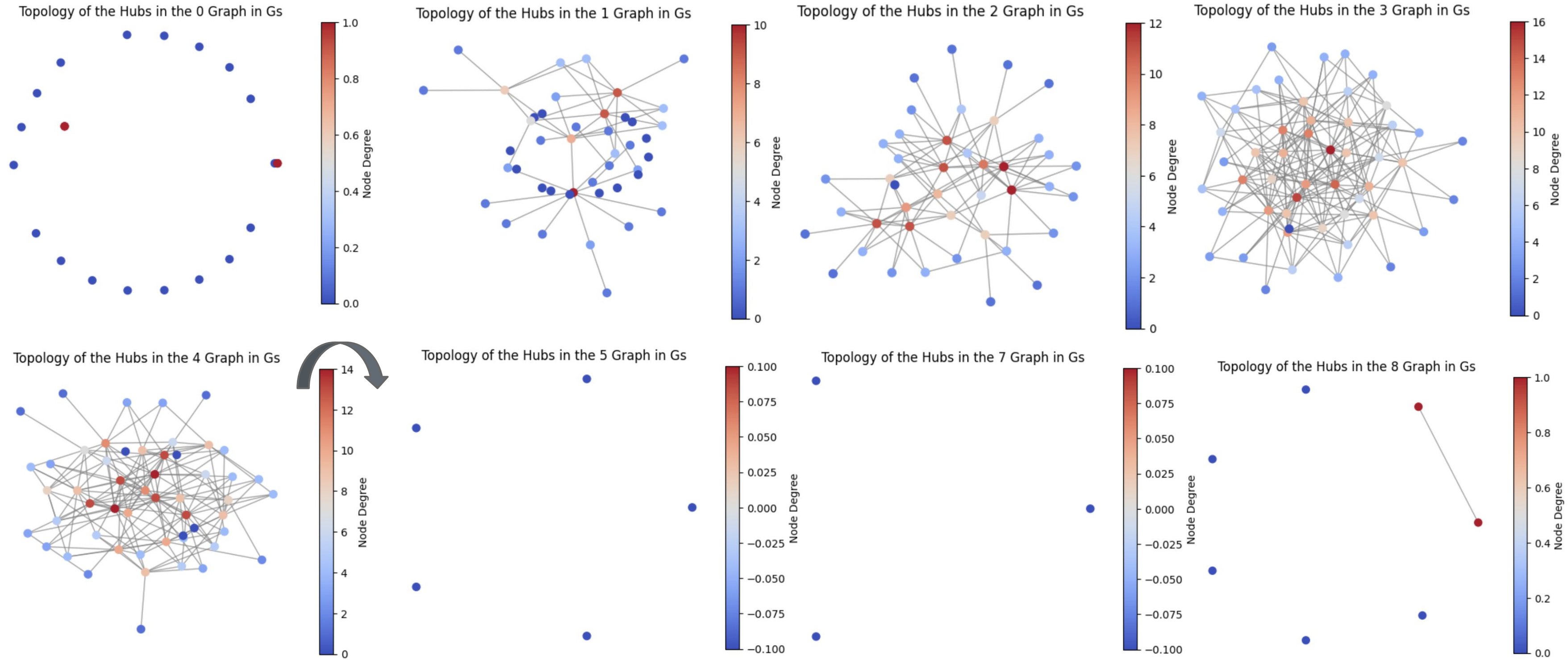


Hubs – Threshold 20



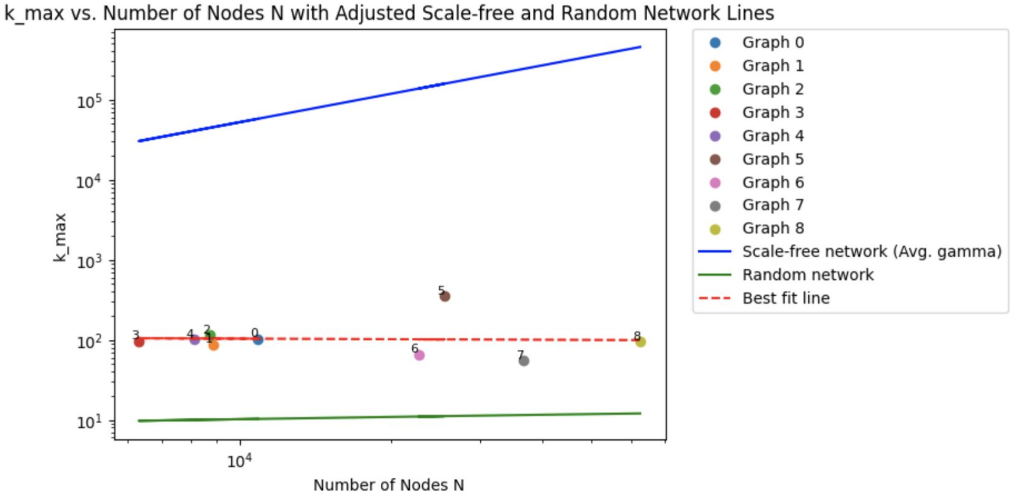
Days: ['04', '05', '06', '08', '09', '24', '25', '30', '31']

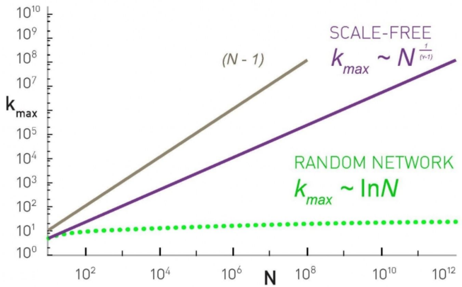
Hubs – Threshold 50



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

Hubs - k_max vs. N





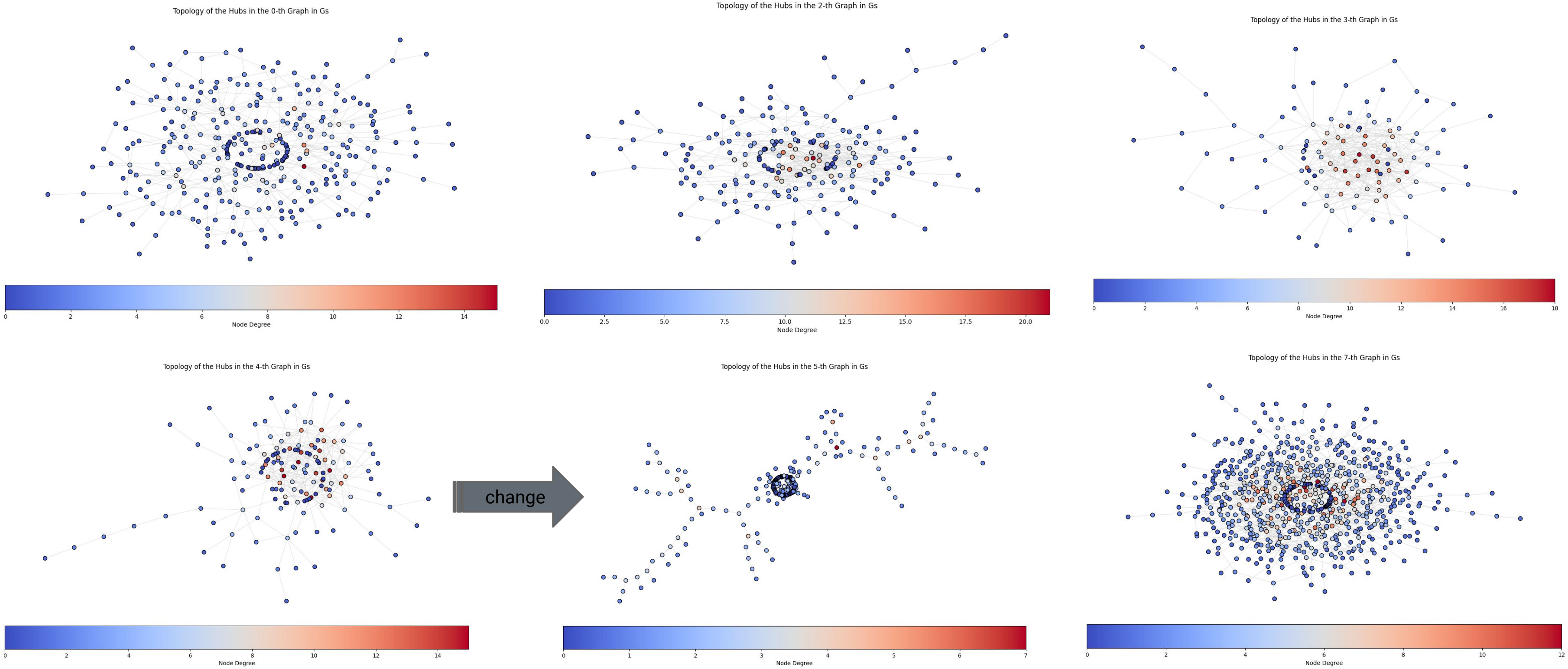
gamma = [1.668, 1.674, 1.673, 1.754, 1.780, 1.986, 1.994, 2.036, 2.063]

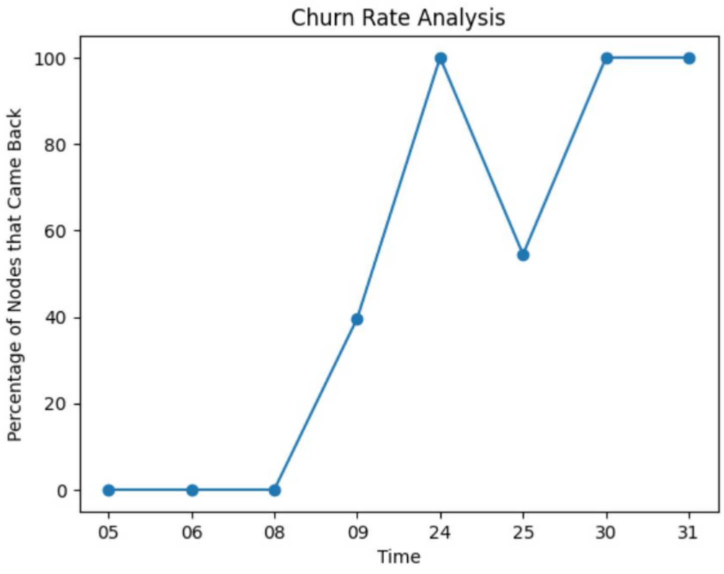
Average gamma = 1.848

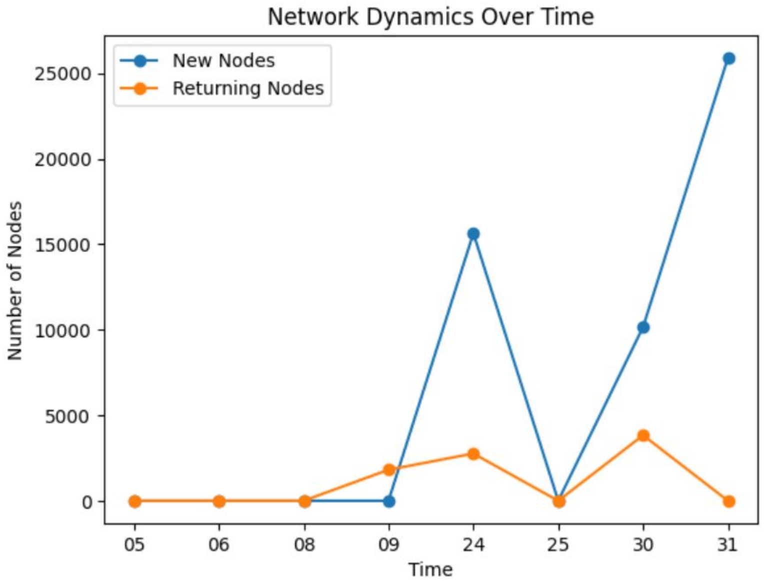
Anomalous Regime (γ **= 2**): or $\gamma = 2$ the degree of the biggest hub grows linearly with the system size, i.e. kmax ~ N. This forces the network into a hub and spoke configuration in which all nodes are close to each other because they all connect to the same central hub. In this regime the average path length does not depend on N.

Future Work

- More consistent dataset in time-series
- Community detection
- SIR simulation on Gnutella network


Topology of the Hubs in the 0-th Graph in Gs
0 2 4 6 8 10 12 14
Node Degree
Topology of the Hubs in the 2-th Graph in Gs
0.0 2.5 5.0 7.5 10.0 12.5 15.0 17.5 20.0
Node Degree
Topology of the Hubs in the 3-th Graph in Gs
0 2 4 6 8 10 12 14 16 18
Node Degree
Topology of the Hubs in the 4-th Graph in Gs
0 2 4 6 8 10 12 14
Node Degree
change
Topology of the Hubs in the 5-th Graph in Gs
0 1 2 3 4 5 6 7
Node Degree
Topology of the Hubs in the 7-th Graph in Gs
0 2 4 6 8 10 12
Node Degree


Churn Rate Analysis
Percentage of Nodes that Came Back
100
80
60
40
20
0
05
06
08
09
24
25
30
31
Time


Network Dynamics Over Time
New Nodes
Returning Nodes
Number of Nodes
25000
20000
15000
10000
5000
0
05
06
08
09
24
25
30
31
Time