Mapping the Gnutella Network

Matei Ripeanu (Univ. of Chicago)
Adriana Iamnitchi (Univ. of Chicago)
Ian Foster (Univ. of Chicago & Argonne National Laboratory)

What are P2P Networks ?

- Networks to share information and resources directly between computers without dedicated servers
- Computers join and leave the network frequently
- Computers in the network may not have permanent IP addresses
- P2P networks are overlay networks. They have their own application level routing mechanisms
- 2 important factors that affect P2P Network's performance are
  - Topology of the overlay network
  - Routing mechanism

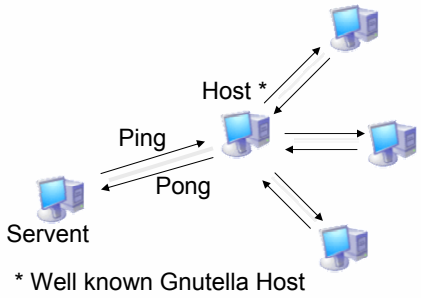
Goal of this Paper

- Study whether Gnutella's overlay network topology maps to the underlying physical Internet infrastructure. (i.e. evaluate the topology mismatch between them if any)
- Evaluate cost & benefits of the Gnutella network
- Investigate possible improvements for scaling the Gnutella network and other similar networks

What is Gnutella ?

- Open, decentralized group membership and search protocol use mainly for file sharing
- Virtual network of hosts that run Gnutella speaking applications
- 4 Goals of the Gnutella protocol
  - Dynamic Operability
  - Performance and Scalability
  - Reliability
  - Anonymity

Gnutella Architecture



* Well known Gnutella Host

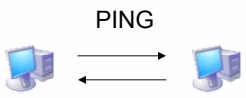
- PING message broadcast to all attached nodes
- PONG messages backtracked to the servent

- Indefinite propagation prevented by setting "TTL" & "Hops-passed"

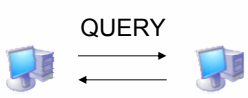
Overlay Network = Servents + TCP Connections
(Routers) (Links)

Gnutella Messages

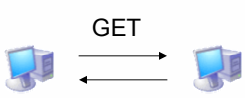
- Gnutella Protocol mainly supports 6 messages



- Broadcast
- Back-Propagated



- Broadcast
- Back-Propagated

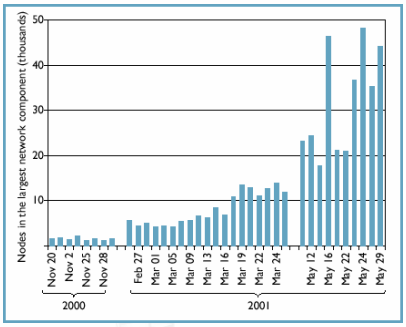


- Node to Node

Crawler

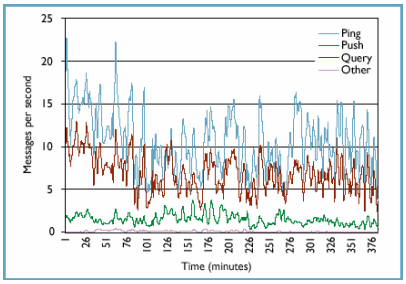
- Authors developed a crawler
- 2 versions - sequential & client server
- Connects to Gnutella nodes present in a predefined list
- Creates TCP connections and PINGs the nodes
- Builds a list of discovered nodes using PONG messages
- List contains node IP Addresses, Port, No. of files shared, total space of shared files
- Authors used the data gathered by the crawler for analysis

Analysis (Growth Trends)



- Study done over period of 7 months
- 40% nodes leave network < 4 hours
- 25% nodes remain alive > 24 hours
- Exponential growth due to low starting user base, initial user curiosity, availability of broadband access, open architecture, wider choice of clients

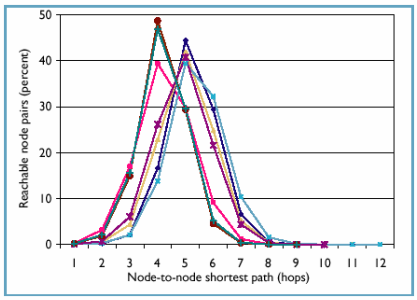
Analysis (Generated Traffic)



- Total Gnutella traffic (excluding file transfers) amounts to 1.7% of total traffic over US Internet backbone
- Serious traffic issues that affect its scalability

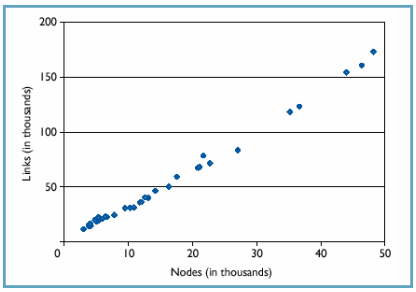
Message Type	Gnutella (Nov 2001)	Gnutella (Jun 2002)
User (QUERY)	36 %	91 %
Group Membership (PING/PONG)	55 %	8 %
Other Non-Standard	9 %	1 %

Analysis (Node - Node Shortest Path)



- Study done by performing 7 crawls of the network
- 95% of any 2 node pairs could exchange messages in 7 hops

Analysis (Average Node Connectivity)

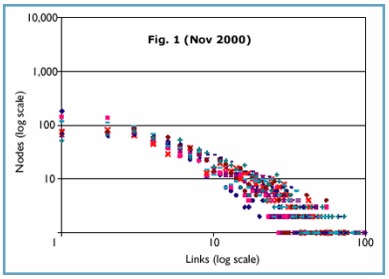


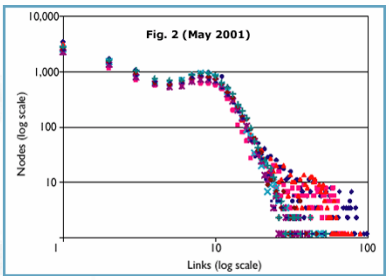
- Average number of connections per node remained constant at about 3.4 as the size of network increased
- Assuming this characteristic to hold true in other cases as well, we could estimate traffic and scalability limits for larger networks as well

Analysis (Connectivity & Reliability)

- Gnutella is a self organizing network. Users decide max. no of connections and nodes decide which other nodes to connect to
- These networks behave like power-law networks which means -
  - Highly stable and resilient, but prone to occasional failure
  - Few nodes with high connectivity and many nodes with low connectivity
  - Extremely robust to random node failure, but vulnerable to planned attacks

.. Analysis (Connectivity & Reliability)



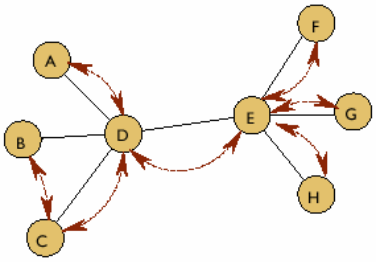


- Fig.1 shows a quasi-linear distribution (characteristic of power law networks)
- Fig.2 does not show characteristic of power law networks for nodes < 10 links. Shows a heavy tailed graph. for Nodes > 10 links
- Reason - Few nodes with high connectivity act like servers & are 50% more likely to be connected to by new node
- Fig.1 - Less dependence on highly connected nodes. More fault tolerant
- Fig.2 - More dependence on highly connected nodes. Reduces the reliability of the network

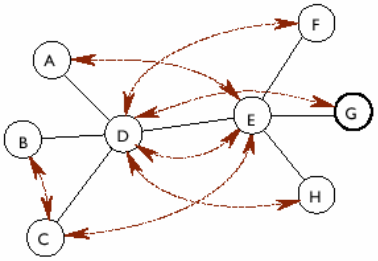
Topology Mapping

- 2 Topologies that we are talking about are
  - Internet topology
  - Gnutella's overlay topology
- Gnutella's overlay topology should preferably perfectly map to the underlying Internet topology
- Mismatch of these two topologies leads to
  - Inefficient use of resources
  - Increases cost for ISPs
  - Limits scalability
- Gnutella's overlay topology does not match the underlying Internet topology. Authors performed 2 high level experiments to highlight this topology mismatch

Topology Mapping (contd.)



Perfect Mapping



Inefficient Mapping

- When Gnutella's overlay network topology closely matches the underlying infrastructure, a broadcast from A involves only one communication over link DE
- When there is a mismatch, the same broadcast involves 6 communications over link DE

Topology Mismatch (Experiments)

- Experiment 1
  - More than 40% of Gnutella nodes lie in the top 10 Autonomous Systems (AS) in the Internet
  - In spite of that, only 2 - 5 % of Gnutella connections link nodes within a single AS
  - Most Gnutella generated traffic crosses AS borders
  - This increases costs and limits scalability

Topology Mismatch (Experiments)

- Experiment 2
  - Assumed domain names express hierarchy Nodes identified with their domain names
  - Calculate entropy of cluster of Gnutella nodes and compare with entropy of a random selection of nodes from across domains
  - Gnutella graph is random if entropy is same. Otherwise it is more ordered
  - Authors find Gnutella clustering entropy to be 8% lower than random clustering entropy

Problems with Gnutella

- Topology Mismatch
- Security – Topology information can be obtained easily and planned DOS attacks can cause harm
- No guarantee you will find desired resource because it can be greater than 8 hops away
- Search results take time since all responses from 8 levels deep have to back propagate
- More bandwidth used since you are a part of the chain. Servents respond & forward other servents requests
- Overload Gnutella network with a flood of bogus search packets. Source cannot be traced because identity is preserved

Scope for Improvement

- Improve security
- Distribution of Gnutella queries follow the Zipf's Law. Exploit this by making use of proxy cache mechanisms
- Query-caching scheme along with grouping of nodes by user interest
- Replace query flooding mechanism with smarter routing mechanisms
- Method adopted by Freenet - data caching at nodes

References

- The Annotated Gnutella Protocol Specification v0.4 (1)
  http://rfc-gnutella.sourceforge.net/developer/stable/index.html

[Primers]

- A Definition of Peer-to-Peer Networking for the Classification of Peer-to-Peer Architectures and Applications – Rudiger Schollmeier
- An Introduction to Peer-to-Peer Computing – David Barkai

[Papers]

- Tracing a large scale Peer to Peer System: an hour in the life of Gnutella – Evangelos P. Markatos
- Free Riding on Gnutella – Eytan Adar and Bernardo A. Huberman
- Zipf's Law and the Internet – Lada A. Adamic, Bernardo A. Huberman

[Websites]

- Knowbuddy's Gnutella FAQ
  http://www.rixsoft.com/Knowbuddy/gnutellafaq.html
- www.howstuffworks.com