Gnutella Crawling
Lessons Learned

David Deschenes and Scott Weber
Department of Computer Science & Engineering
Lehigh University

April 30, 2003
Outline

- Gnutella Protocol Changes
- Crawling Concepts
- Crawler Architecture
- Comparison of Protocol Versions
- Characterizing Valuable Hosts
- Exploring Ping Depths
- Looking at Time Effects
- Summary
- Future Work
Gnutella Protocol Changes

- Two Versions: 0.4 and 0.6
- Version 0.6 introduced in late 2001
- Version 0.6 supports protocol extensions by allowing exchange of header information
- The X-Try header may have implications for future Gnutella crawlers
Crawling Concepts

• Similar to WWW document crawling
• Must make use of host discovery mechanisms (ping/pong messages, X-Try headers)
• Pong messages encode network links
• Will not ever know if entire network has been mapped
Comparison of Protocol Versions

- Simultaneous 0.4 and 0.6 crawls run eight times daily April 10-16
- Each crawl used an identical start set of about 50,000 hosts
- Goals
  - Determine extent of deployment and impact of version 0.6
  - Determine usefulness of each protocol version in mapping the network
Host Identifications Per Visit

![Graph showing average number of host identifications per visit by time of day for Version 0.4 Total, Version 0.6 by Pong, and Version 0.6 Total Only.](image-url)
Measured Network Sizes

![Graph showing network size over time](image-url)
Characterizing Valuable Hosts

- Analysis of data obtained from protocol version comparison experiment
- Goals
  - Determine how to identify the hosts that will give us the best topology data
  - Determine how to order those hosts
Sessions Breakdown

- **Discovered**
  - 60.73%
  - 39.27%

- **Known**
  - 75.58%
  - 24.42%

Legend:
- Discovered
- Known
- Identified
- Unidentified
Value of Identification Types

Pie Chart:
- 67.68% Pong Only
- 9.84% X-Try Only
- 22.48% Pong & X-Try

Bar Chart:
- Pong Only: 0.0000, 0.0250, 0.0500, 0.0750, 0.1000, 0.1250, 0.1500, 0.1750, 0.2000, 0.2250, 0.2500
- X-Try Only: 0.0000, 0.0250, 0.0500, 0.0750, 0.1000, 0.1250, 0.1500, 0.1750, 0.2000, 0.2250, 0.2500
- Pong & X-Try: 0.0000, 0.0250, 0.0500, 0.0750, 0.1000, 0.1250, 0.1500, 0.1750, 0.2000, 0.2250, 0.2500

Bar chart legend:
- Blue: Pong Only
- Red: X-Try Only
- Yellow: Pong & X-Try

% of Hosts Providing Sessions
Exploring Ping Depths

- Ten crawls run daily April 9-23
- Each crawl used a different TTL for the ping message sent
- Goals
  - Determine which ping depth produces the best topology data
  - Examine change in efficiency of crawl as ping depth varies
Pongs by Hops

![Graph showing the average number of Pongs by Pong Hops and Ping Depth.](image-url)
Ping Depth Efficiency

![Chart showing Ping Depth Efficiency with varying average number of hosts for each depth level.]
Looking at Time Effects

• Analysis of data obtained from protocol version comparison experiment

• Goals
  – Identify diurnal patterns in network activity
  – Contrast weekday with weekend network activity
Fluctuation in Network Activity

% Deviation from Avg.
Weekend Gains

![Weekend Gains Bar Chart]

- Weekend Discoveries: 26%
- Weekend Pongs: 22%
- Weekend Unique Pongs: 22%
- Weekend Sessions: 20%
- Weekend TCP Conns: 16%
- Weekend Measured Size: 18%
- Weekend Session Length: 4%

% Gain Over Weekday
Summary

- Version 0.4 is still active, but cannot be used solely for an effective crawl
- Proper ordering of hosts may yield more efficient crawls
- Choice of ping depth depends on goal of crawl
  - Depth 7 yields good balance between efficiency and quality
- The network is not stable on long or short timescales
  - Must crawl often to remove statistical effects
Future Work

• Modify crawler to take into account information gathered in this experiment
• Study how the Gnutella network changes over time
• Compare complete network measurements to those of LimeWire
Questions?