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

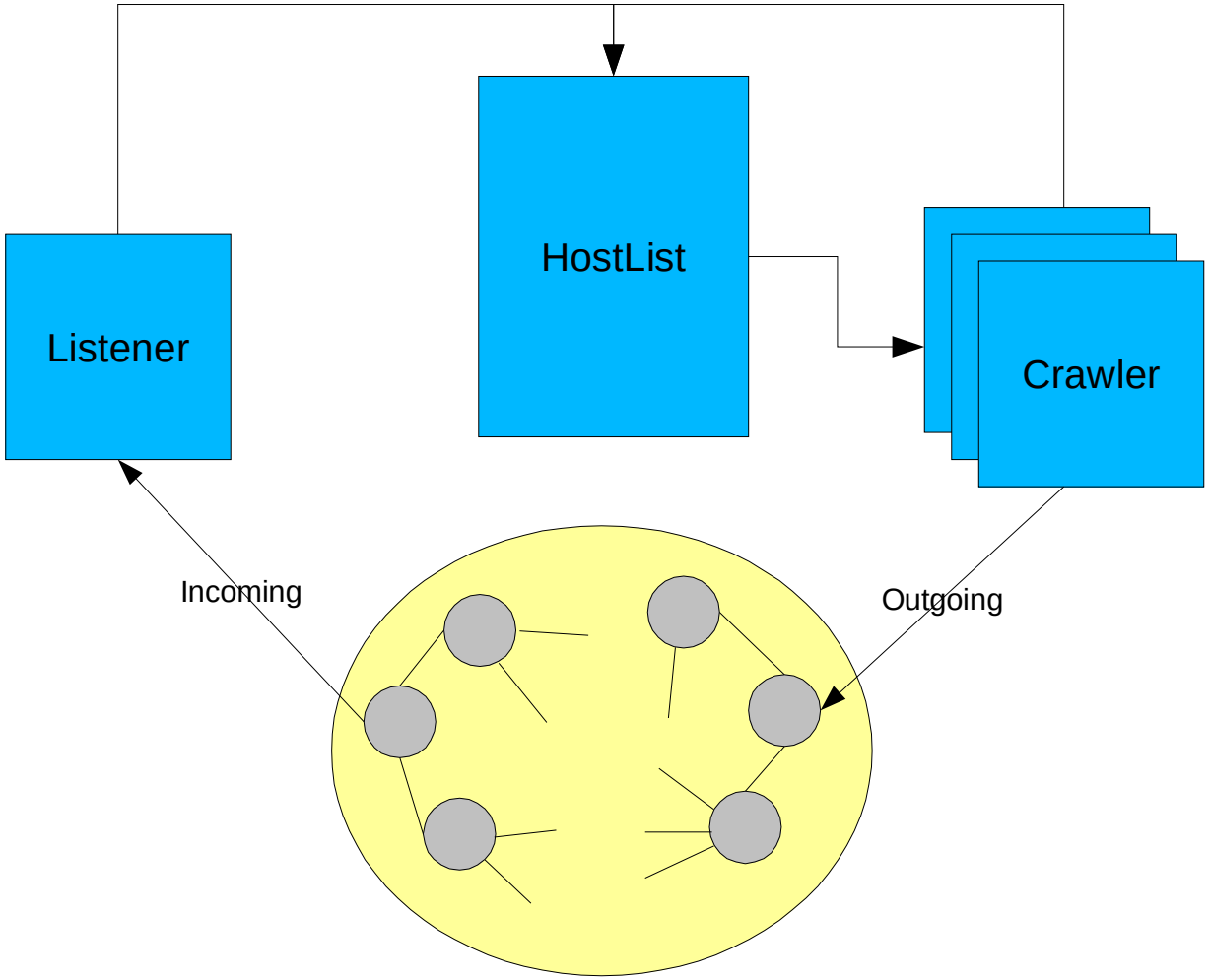
Gnutella Protocol Changes

- 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

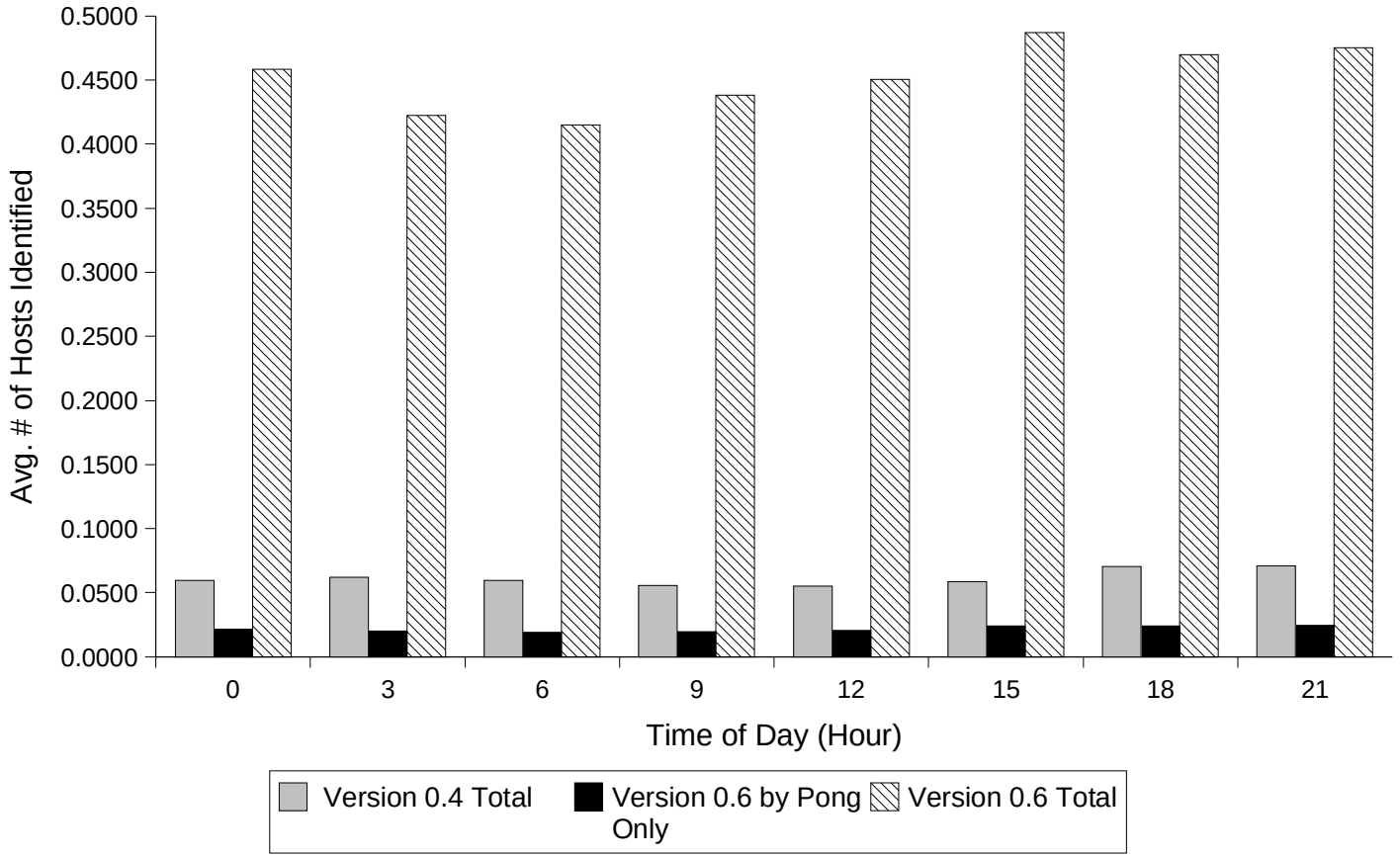
Crawler Architecture



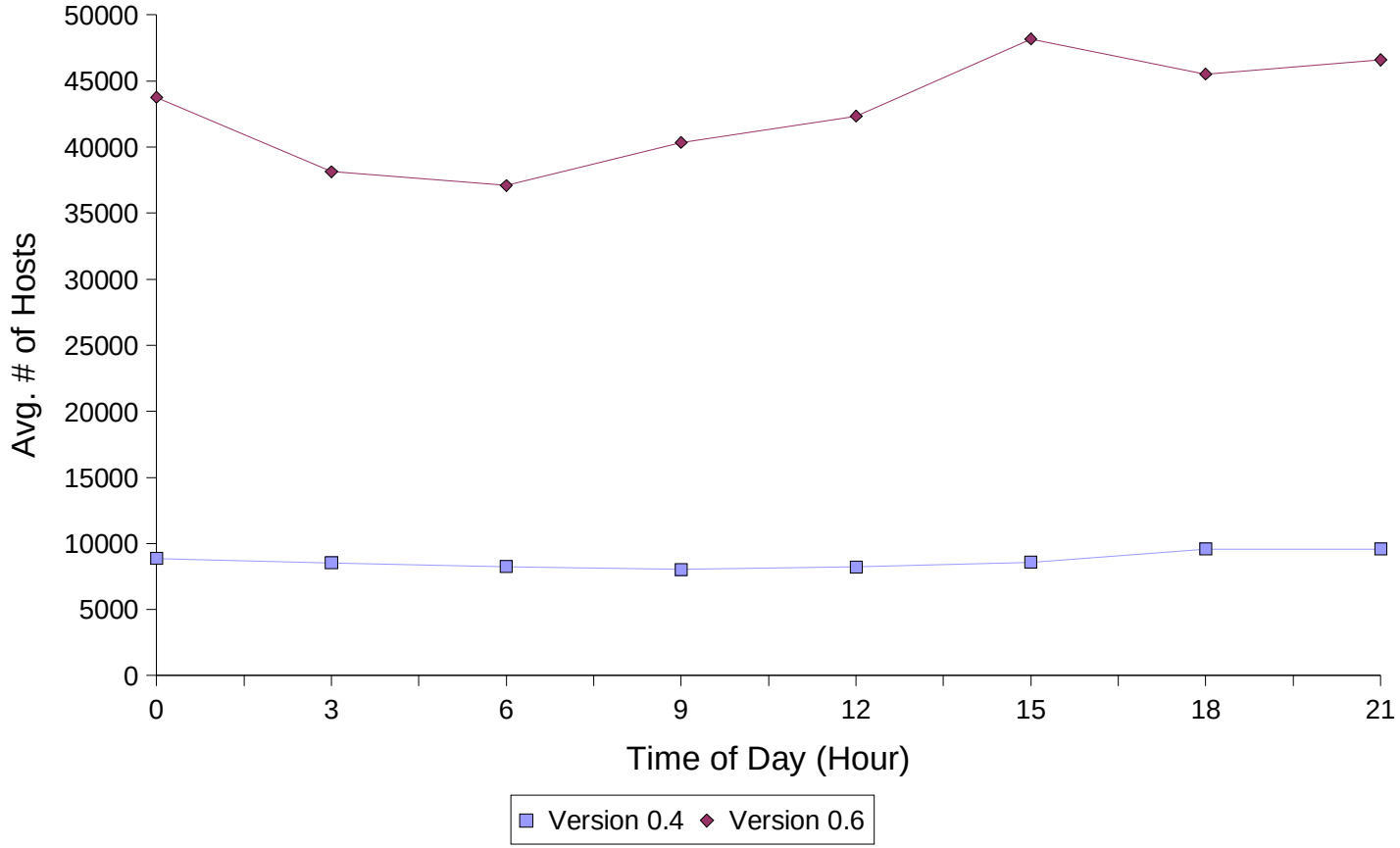
Comparison of Protocol Versions

- 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



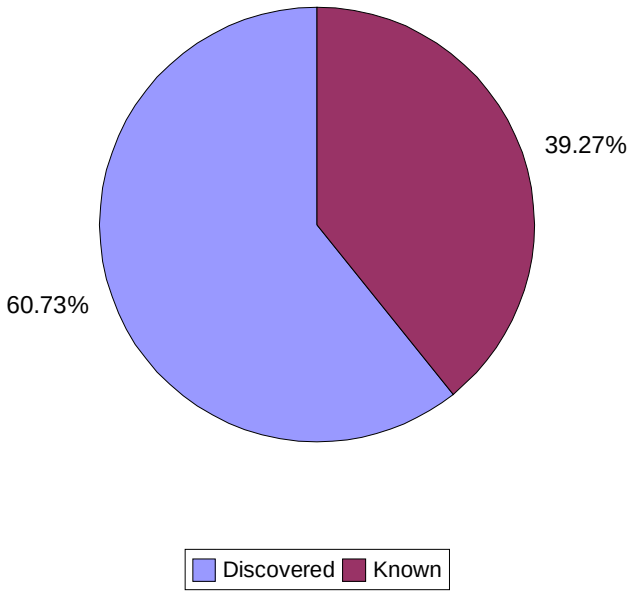
Measured Network Sizes

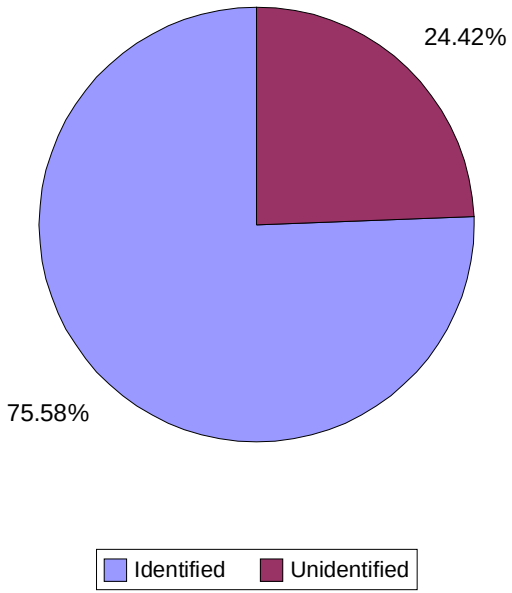


Characterizing Valuable Hosts

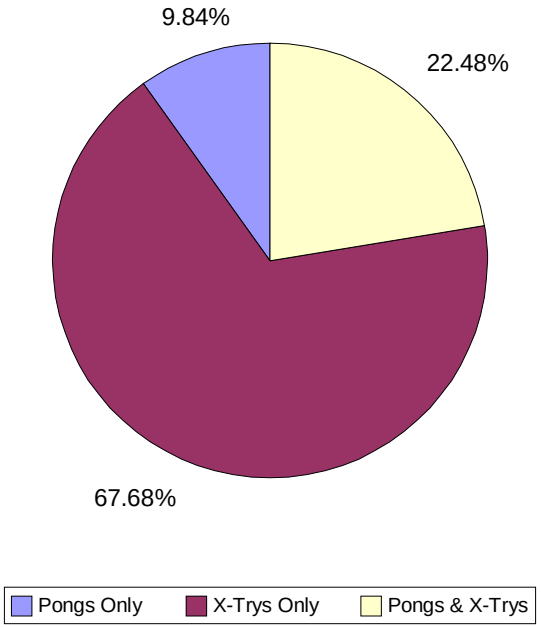
- Analysis of data obtained from protocol version comparison experiment
- Goals
  - Determine how to identify the hosts that will give us the best topology data
  - Determine how to order those hosts

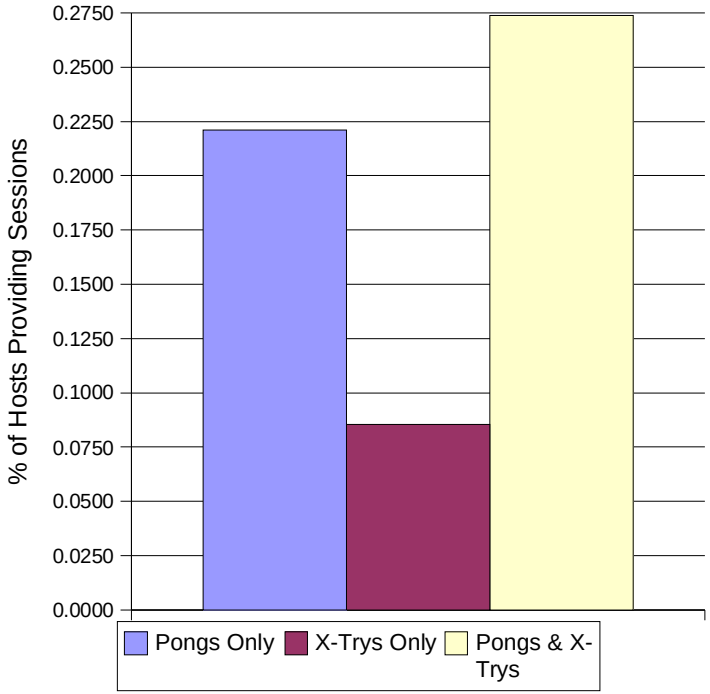
Sessions Breakdown





Value of Identification Types

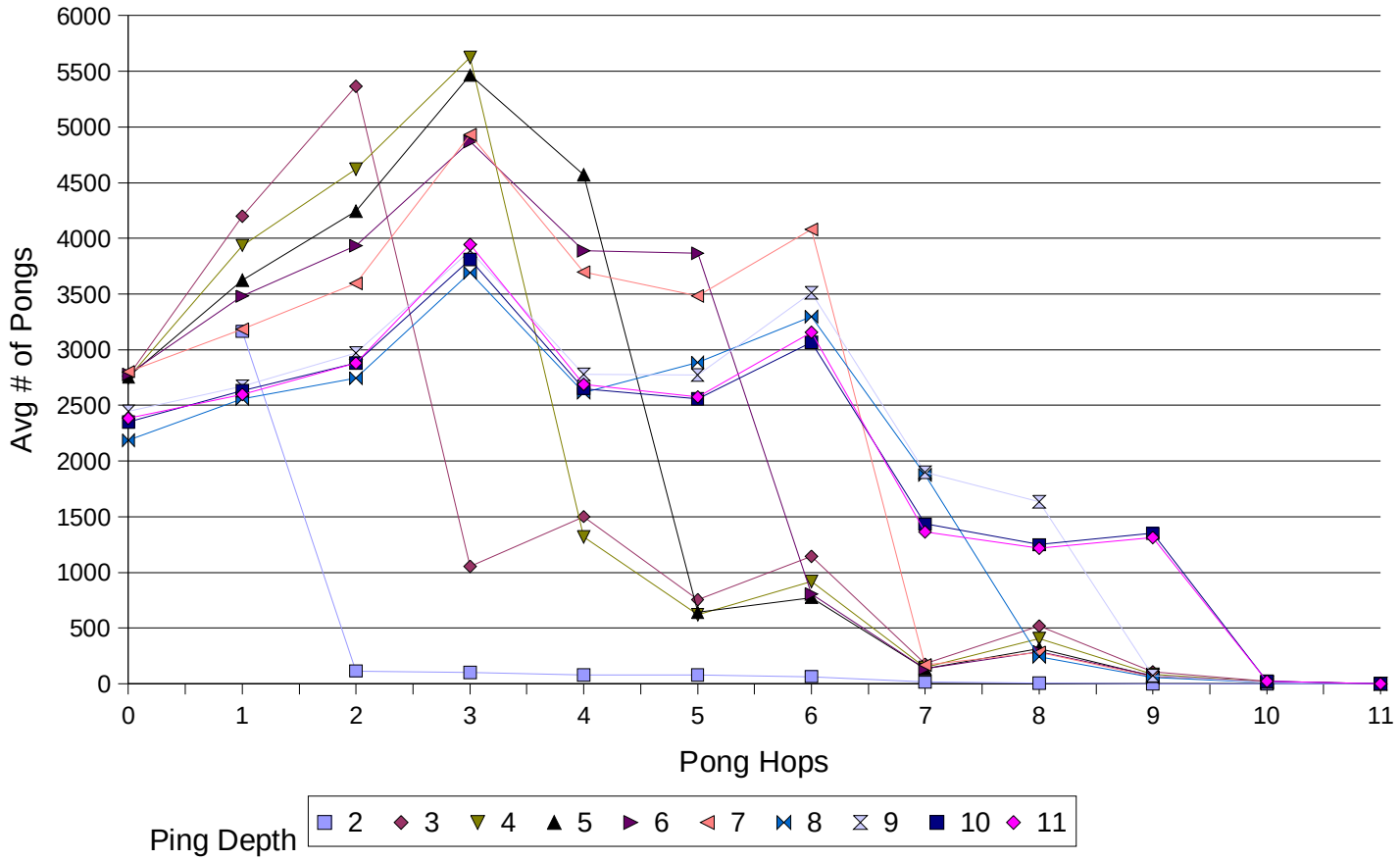




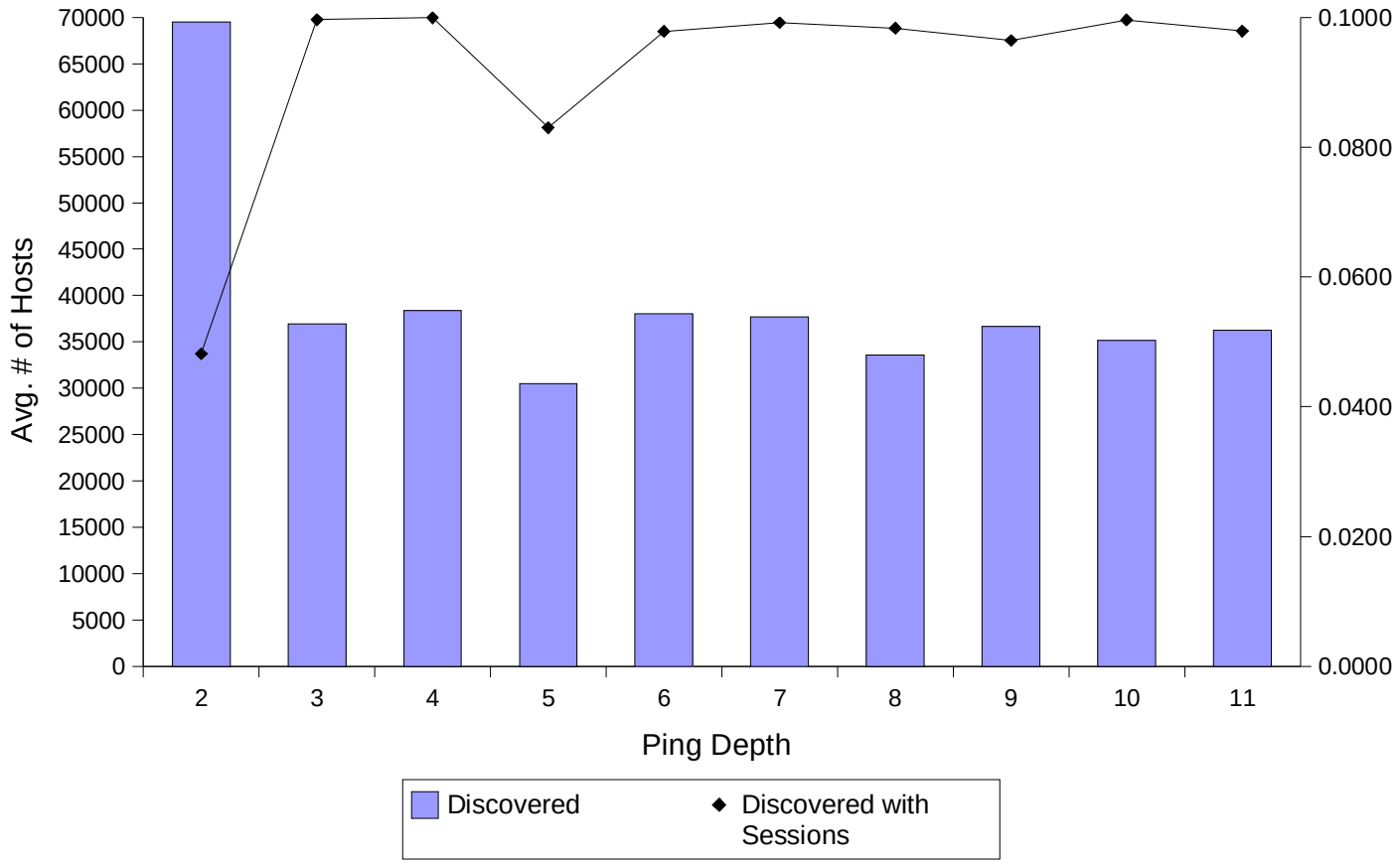
Exploring Ping Depths

- 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 efficieny of crawl as ping depth varies

Pongs by Hops



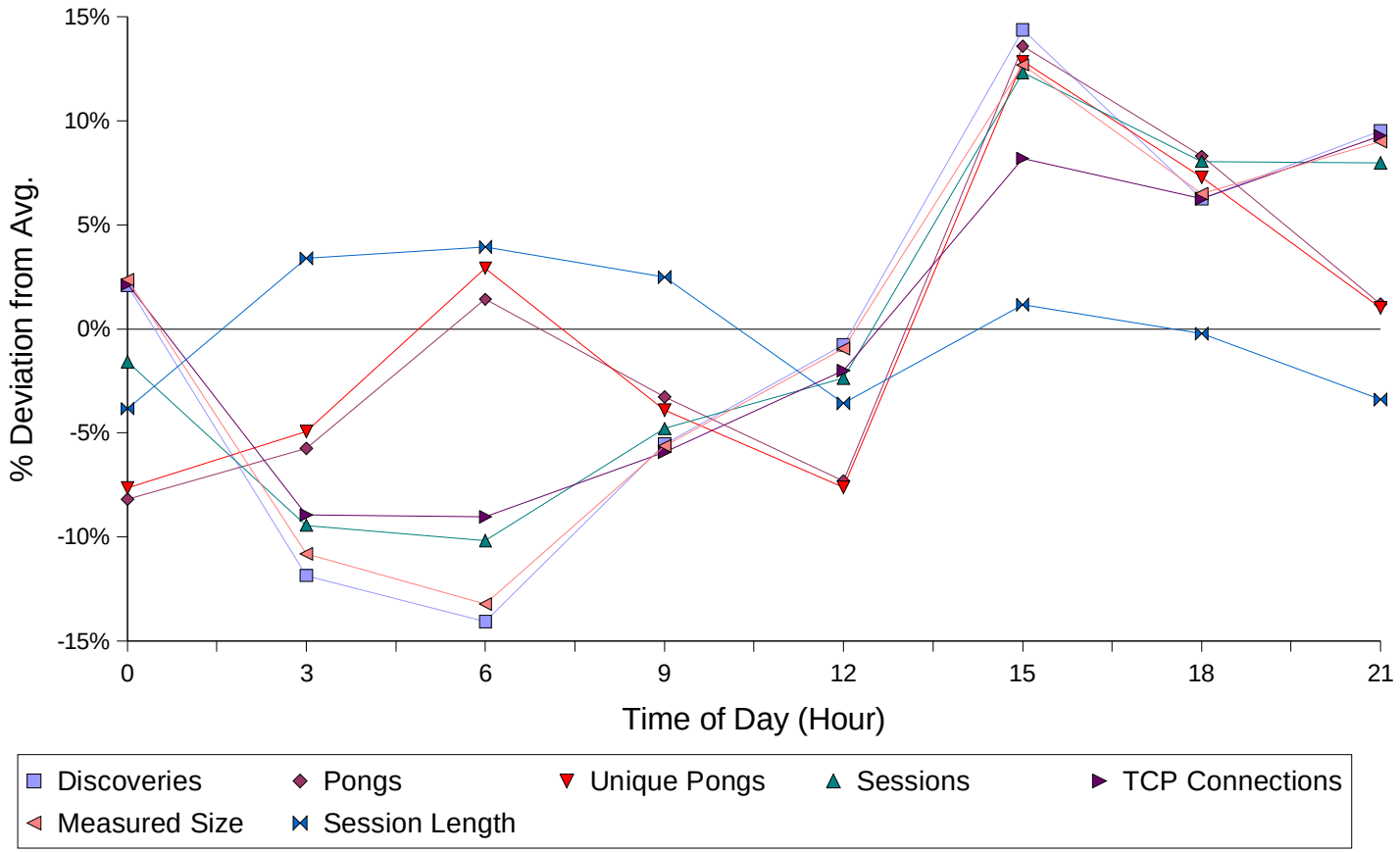
Ping Depth Efficiency



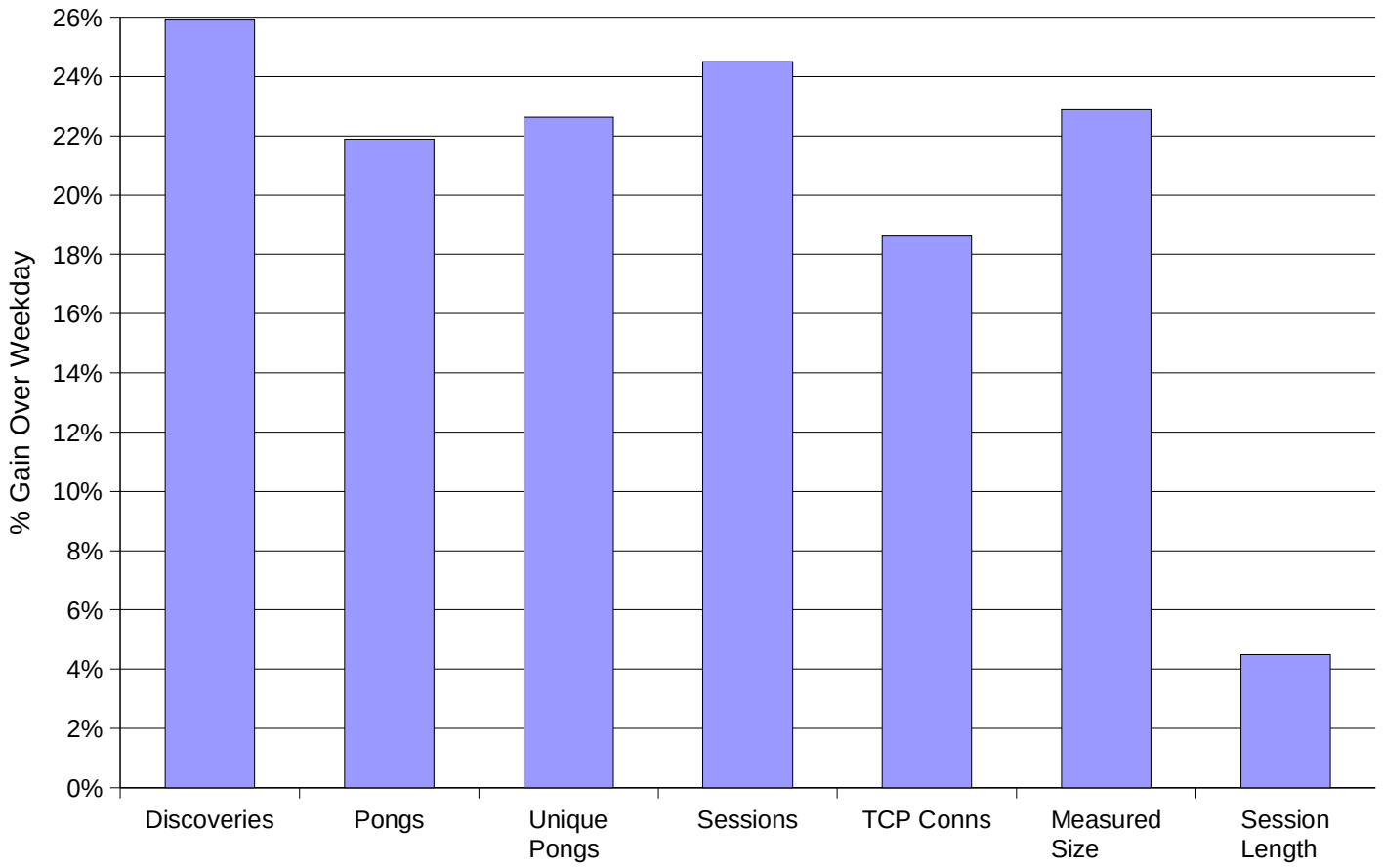
Looking at Time Effects

- 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



Weekend Gains



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?