

Capturing Accurate Snapshots of the Gnutella Network

Daniel Stutzbach

Reza Rejaie

Department of Computer Science

University of Oregon

Global Internet Symposium

Miami, FL

March 19th, 2005

Motivation

- There are two key aspects of P2P systems to characterize:
 - [The overlay topology](#)
 - The overlay traffic
- Characterizing P2P applications is important for:
 - Understanding their performance limitations
 - Examining their impact on the network
 - Simulating new designs
- Characterizing the overlay topology requires [accurate](#) global snapshots of the P2P overlay.
- Peer-to-Peer (P2P) applications are growing in popularity.
 - Several million simultaneous users
 - Significant fraction of network traffic

Characterizing Overlay Topologies

- Snapshots are graphs of peers as vertices and connections as edges.
- A crawler walks the overlay topology to capture snapshots.
- Individual snapshots reveal graph-related properties of the overlay.
- Consecutive snapshots reveal the dynamics of the overlay.
- Characterization quality depends on the accuracy and granularity of captured snapshots.

Challenges in Capturing Accurate Snapshots

- Large and rapidly changing overlay topology
 - Churn: Peers join and leave the overlay.
 - Connections among peers change.
 - Snapshot accuracy is primarily determined by crawling speed.
 - Unreachable peers can affect snapshot accuracy
 - A non-negligible portion of peers are NATed, departed, or overloaded.
 - Each group has a different effect on snapshot accuracy.
- *Capturing accurate snapshot of large P2P system is difficult!*

Related Work on the P2P Topologies

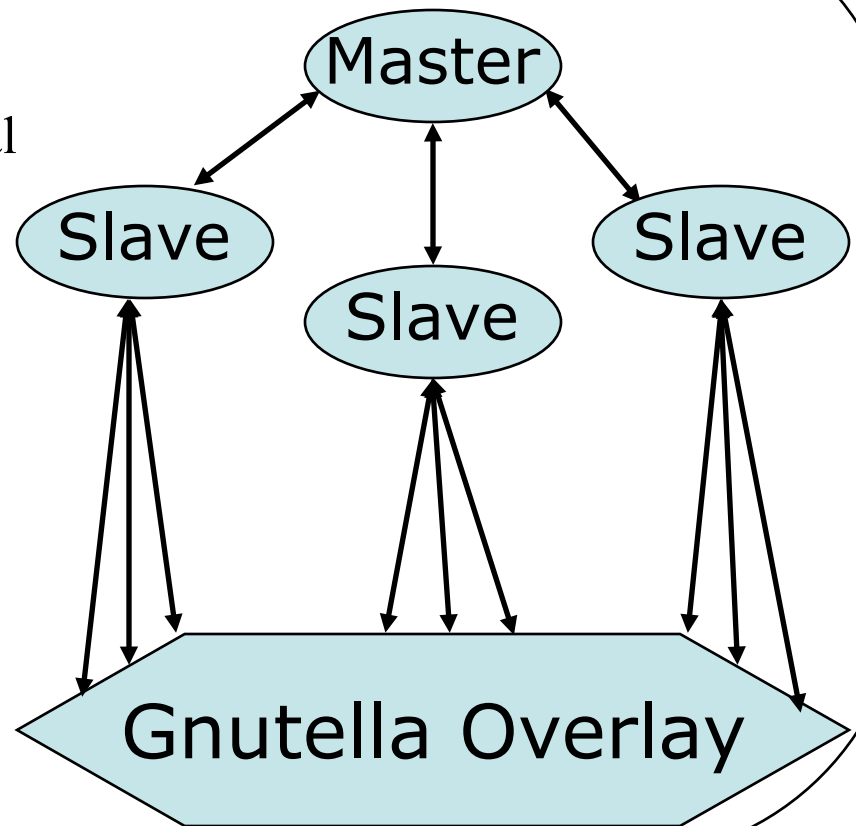
- Capturing complete snapshots
 - 133 peers/min, capturing 8,000 peers in 1 hour [Clip2 00]
 - 250 peers/min, capturing 30,000 peers in 2+ hours [Ripeanu 02]
 - However, many peers are only present for just a few minutes!
- Capturing partial snapshots
 - 2,500 peers/min, capturing 10,000 peers in 2 minutes [Sarioiu 02]
 - Partial snapshots are less distorted but may be unrepresentative.
 - For some types of analysis, the whole graph is needed.
- These studies are outdated because
 - The network size has significantly increased (30k to 1.3 million).
 - Gnutella shifted to a two-tier architecture.
 - No measurement studies of other two-tier topologies.

The Gnutella Cruiser

- Increased crawl speed using:
 - Parallel crawling
 - The two-tier architecture of the overlay
 - Identifying status of unreachable peers
 - Quantifying/estimating their impact on snapshot accuracy.
- *Cruiser can capture the one million node network in around 7 minutes, or 140,000 peers/minute.*

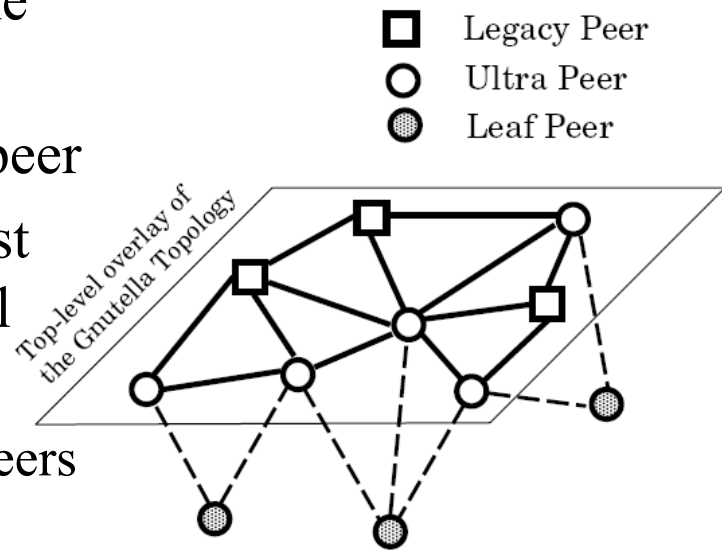
Parallel Crawling

- Cruiser probes hundreds of nodes in parallel.
 - Major speedup vs. sequential crawling
 - The degree of parallelism is dynamically adapted.
- Cruiser uses a master-slave architecture.
 - Further improves performance.
 - Allows for load distribution



Leveraging Two-Tier Topology

- All connections include at least one reachable top-level peer.
 - Each leaf connects to an ultrapeer
 - Firewallled top-level peers must peer with a reachable top-level peer.
 - Therefore, crawl only top-level peers
- This yields an 85% savings.

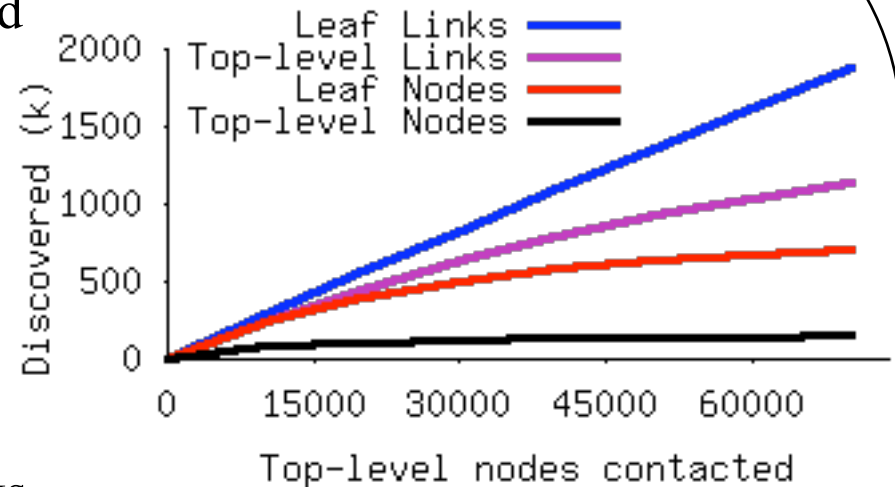


Performance Evaluation

- Evaluating the accuracy of a crawler is difficult, as there are no perfect reference snapshots.
- We explore three aspects of crawler performance:
 - Completeness of snapshots
 - Effects of crawl duration
 - Impact of snapshot accuracy on conducted characterization

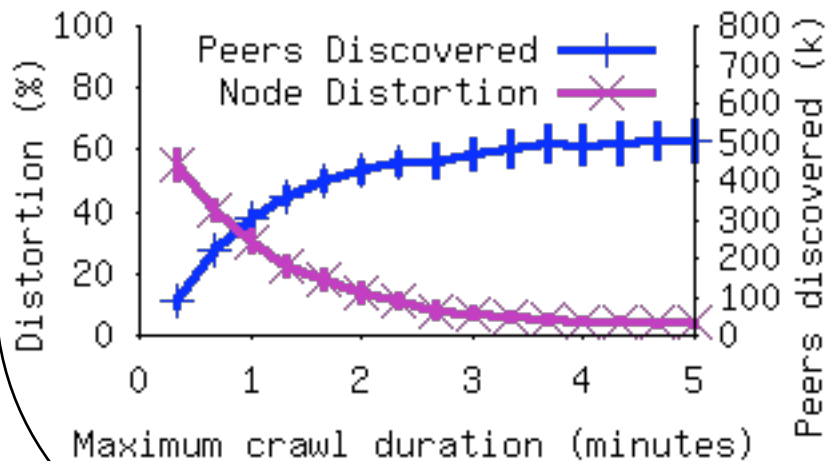
Completeness of Snapshots

- Cruiser terminates after it attempts to contact all discovered top-level peers.
- The incremental value of contacting more peers indicates snapshots are reasonably complete.
 - Top-level peers are discovered quickly.
 - Leaf nodes are top-level links are well-discovered by the end of the crawl.
 - Leaf links necessarily grow linearly.

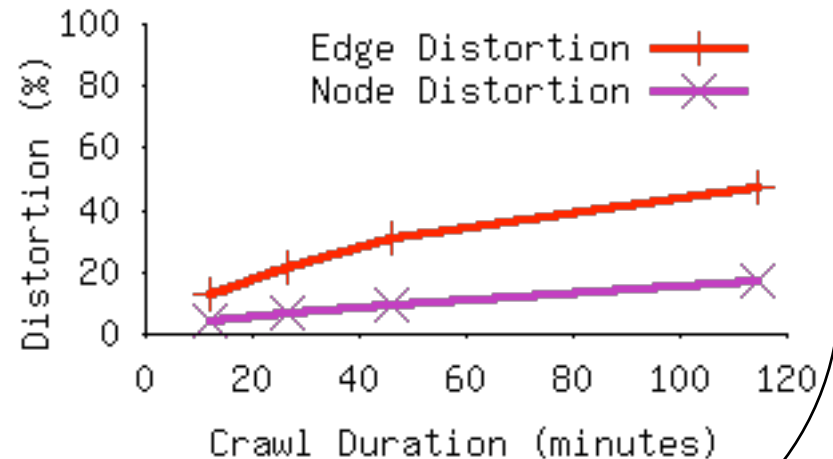


Effects of Crawl Duration

- Comparing back-to-back snapshots estimates distortion.
- Truncated crawls have higher granularity but are less complete.
- Slower crawls are more distorted.



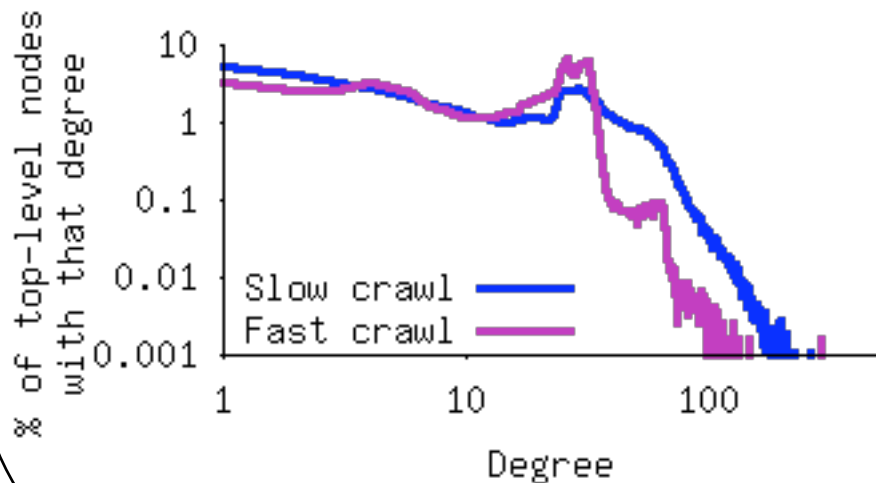
Truncated Crawls



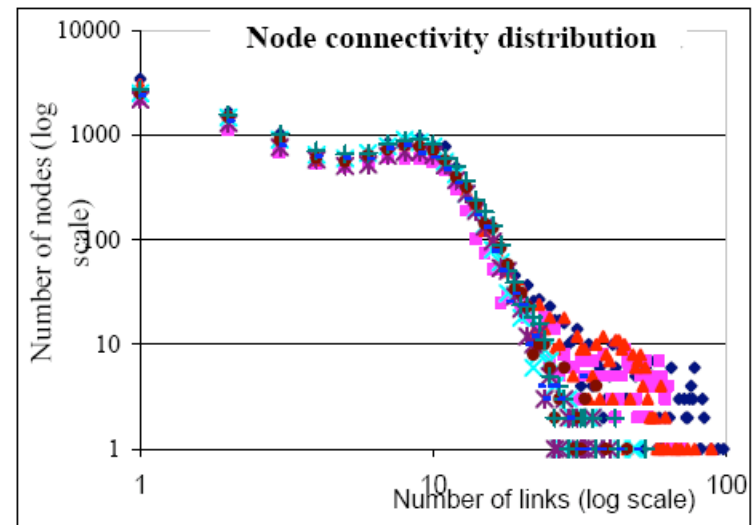
Slow Crawls

Impact of Snapshot Accuracy

- We examine the degree distribution.
- A too-slow crawl has a power-law tail artifact.



Cruiser



[Ripeanu02]

Conclusions

- Cruiser crawls the Gnutella network at 140,000 peers/minute.
- It captures accurate snapshots of the overlay topology.
- We are currently working on measuring and characterizing:
 - Graph properties of the overlay
 - Peer churn
 - Overlay dynamics
- We are also extending Cruiser to other P2P networks.
- Our goal is to provide a complete characterization of P2P networks, i.e., everything needed for a P2P overlay generator.

Thanks!

- Questions?