Recent Problems in Peer-to-peer Content Retrieval

Brian Neil Levine
Dept. of Computer Science
UMass Amherst

The work by BNL presented here was supported in part by National Science Foundation awards ANI-033055 and EIA-0080199.

Motivation

- Peer-to-peer content sharing is one of the largest portions of traffic on the network.
- Illegal (*gnutella, kazaa*) or not (*Apple iTunes*), understanding the characteristics of such traffic is important.
- This talk:
  - Overview of research in p2p traffic measurement.
  - How such measurements can affect p2p design.

Existing P2P Measurements

- *Ripeanu et al.* – Gnutella topology does not match underlying network topology.
- *Markatos* – A simple, query caching scheme can reduce query traffic by a factor of two.
- *Saroiu et al.* – Gnutella bandwidth, latency, and node availability over a 60-hour period.
- *Adar and Huberman* – A free-rider study, using Gnutella's QueryHit messages to infer peer downloads.
- *Chu, Labonte, Levine (2002)* – Measurements of Napster and Gnutella file popularity and session lengths
- *Chu, Labonte, Levine (2003)*– Measurements of all transfers and most libraries in a large p2p system (openNap); evaluation of Chord

Existing P2P Commercial Apps

- Napster-clones
- Gnutella-clones
- Kazaa
  - Hierarchical Super Nodes index (200-300?) other nodes. Peers ping 5 and take best one. Queries can be forwarded to other supernodes
- Win MX
- Apple iTunes

Searching for Content

- Distributed Hash Tables
  - CAN, Chord, Pastry

  → Distribute the index; Update pointers to content

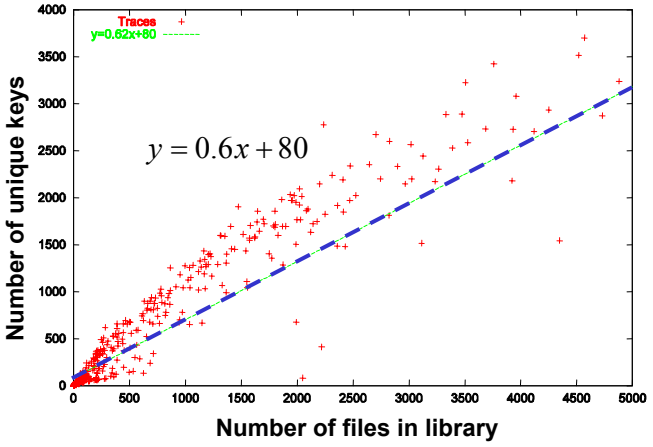
- Gnutella-like search over
  - Small-world networks
  - Power-law degree networks
  - Random graphs

  → Return results only on the content you have stored; Make it easy for searches to traverse the graph; Update the graph; group similar nodes together

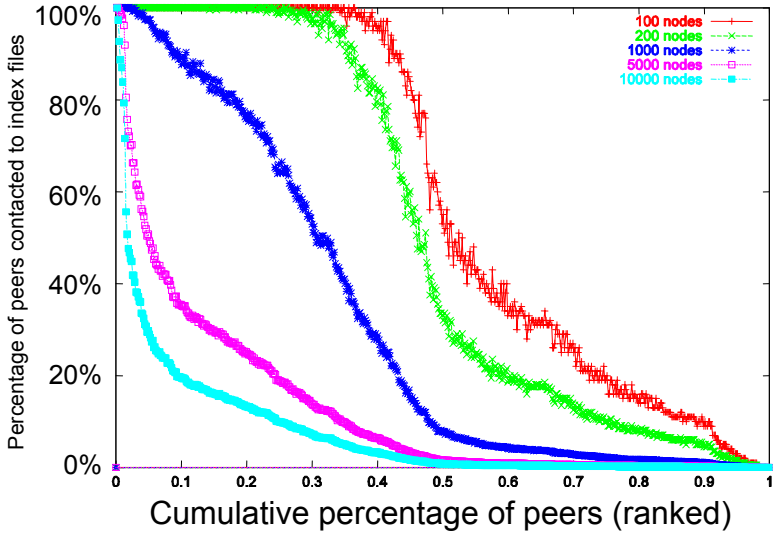
Issues with DHTs

- DHTs work great when
- Libraries of content:

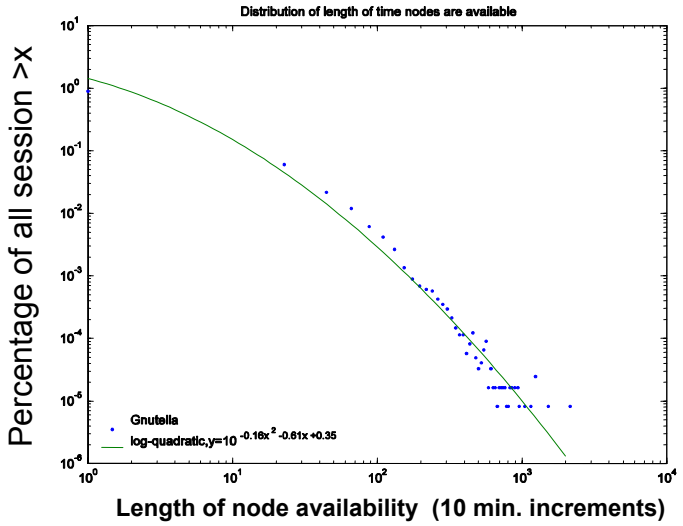
Unique terms in collections of mp3 (file names only!)



Perc. of DHT contacted to index or delete files

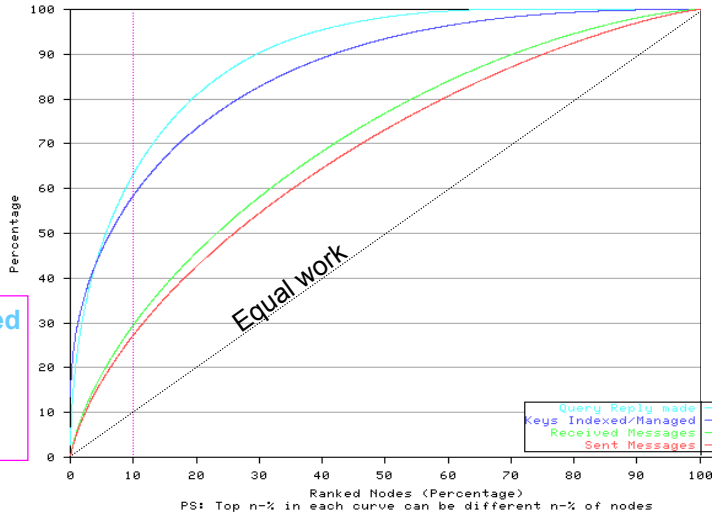


Session Lengths (gnutella)

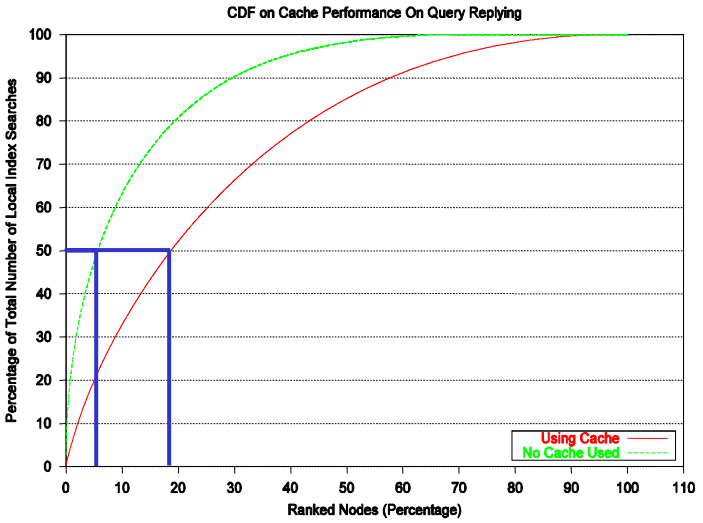


Balance of work in Chord

Queries Resolved
Keys indexed
Msgs sent
Msgs received



Biased workload of resolving searches (Chord)



Open Issues

- Measurement studies have revealed the skewed distributions of p2p systems.
  - Can these be modeled?
- DHTs are limited in their application to content sharing.
  - Work well for single-key systems
- Stronger efforts are needed to match research designs to real characteristics of systems.

- Thanks to Jacky Chu and Kevin Labonte.