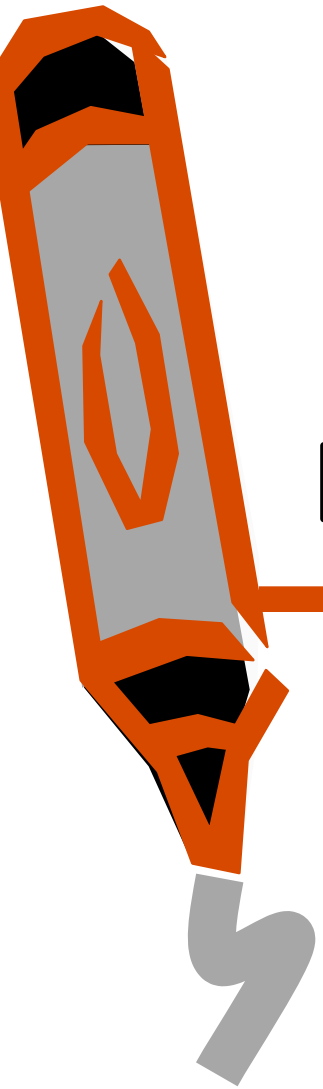
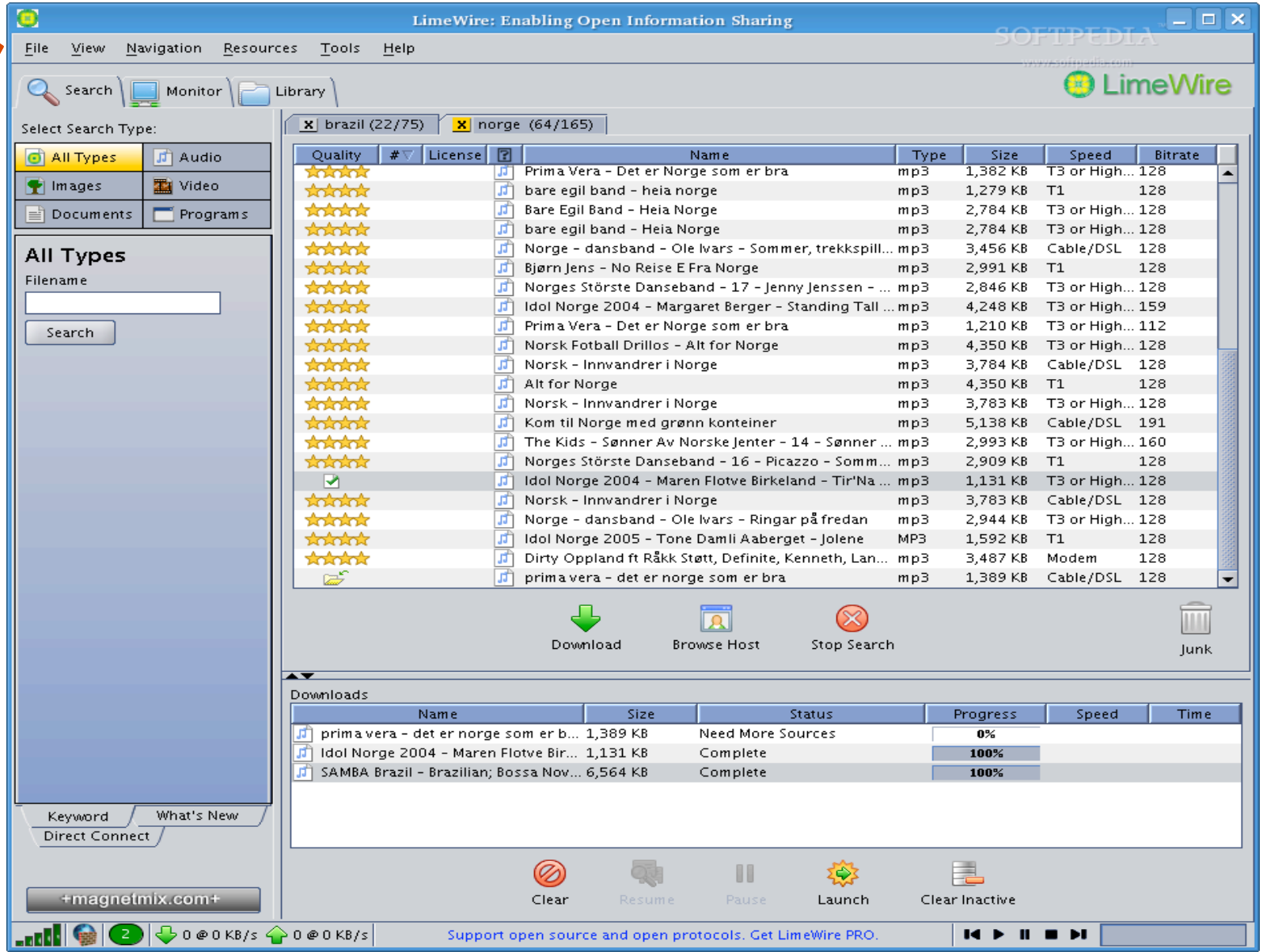


Peer to Peer networks

some slides courtesy James Allan@umass

P2P example


LimeWire: Enabling Open Information Sharing
File View Navigation Resources Tools Help
Search Monitor Library
LimeWire
Select Search Type:
All Types Audio
Images Video
Documents Programs
All Types
Filename
Search
brazil (22/75)
norge (64/165)
Quality # License Name Type Size Speed Bitrate
Prima Vera - Det er Norge som er bra mp3 1,382 KB T3 or High... 128
bare egil band - heia norge mp3 1,279 KB T1 128
Bare Egil Band - Heia Norge mp3 2,784 KB T3 or High... 128
bare egil band - Heia Norge mp3 2,784 KB T3 or High... 128
Norge - dansband - Ole Ivars - Sommer, trekkspill... mp3 3,456 KB Cable/DSL 128
Bjørn Jens - No Reise E Fra Norge mp3 2,991 KB T1 128
Norges Største Danseband - 17 - Jenny Jenssen - ... mp3 2,846 KB T3 or High... 128
Idol Norge 2004 - Margaret Berger - Standing Tall ... mp3 4,248 KB T3 or High... 159
Prima Vera - Det er Norge som er bra mp3 1,210 KB T3 or High... 112
Norsk Fotball Drillos - Alt for Norge mp3 4,350 KB T3 or High... 128
Norsk - Innvandrer i Norge mp3 3,784 KB Cable/DSL 128
Alt for Norge mp3 4,350 KB T1 128
Norsk - Innvandrer i Norge mp3 3,783 KB T3 or High... 128
Kom til Norge med grønn konteiner mp3 5,138 KB Cable/DSL 191
The Kids - Sønner Av Norske Jenter - 14 - Sønner ... mp3 2,993 KB T3 or High... 160
Norges Største Danseband - 16 - Picazzo - Somm... mp3 2,909 KB T1 128
Idol Norge 2004 - Maren Flotve Birkeland - Tir'Na ... mp3 1,131 KB T3 or High... 128
Norsk - Innvandrer i Norge mp3 3,783 KB Cable/DSL 128
Norge - dansband - Ole Ivars - Ringar på fredan mp3 2,944 KB T3 or High... 128
Idol Norge 2005 - Tone Damli Aaberget - Jolene MP3 1,592 KB T1 128
Dirty Oppland ft Råkk Støtt, Definite, Kenneth, Lan... mp3 3,487 KB Modem 128
prima vera - det er norge som er bra mp3 1,389 KB Cable/DSL 128
Download
Browse Host
Stop Search
Junk
Downloads
Name Size Status Progress Speed Time
prima vera - det er norge som er b... 1,389 KB Need More Sources 0%
Idol Norge 2004 - Maren Flotve Bir... 1,131 KB Complete 100%
SAMBA Brazil - Brazilian; Bossa Nov... 6,564 KB Complete 100%
Clear
Resume
Pause
Launch
Clear Inactive
Keyword What's New
Direct Connect
+magnetmix.com+
0 @ 0 KB/s 0 @ 0 KB/s
Support open source and open protocols. Get LimeWire PRO.

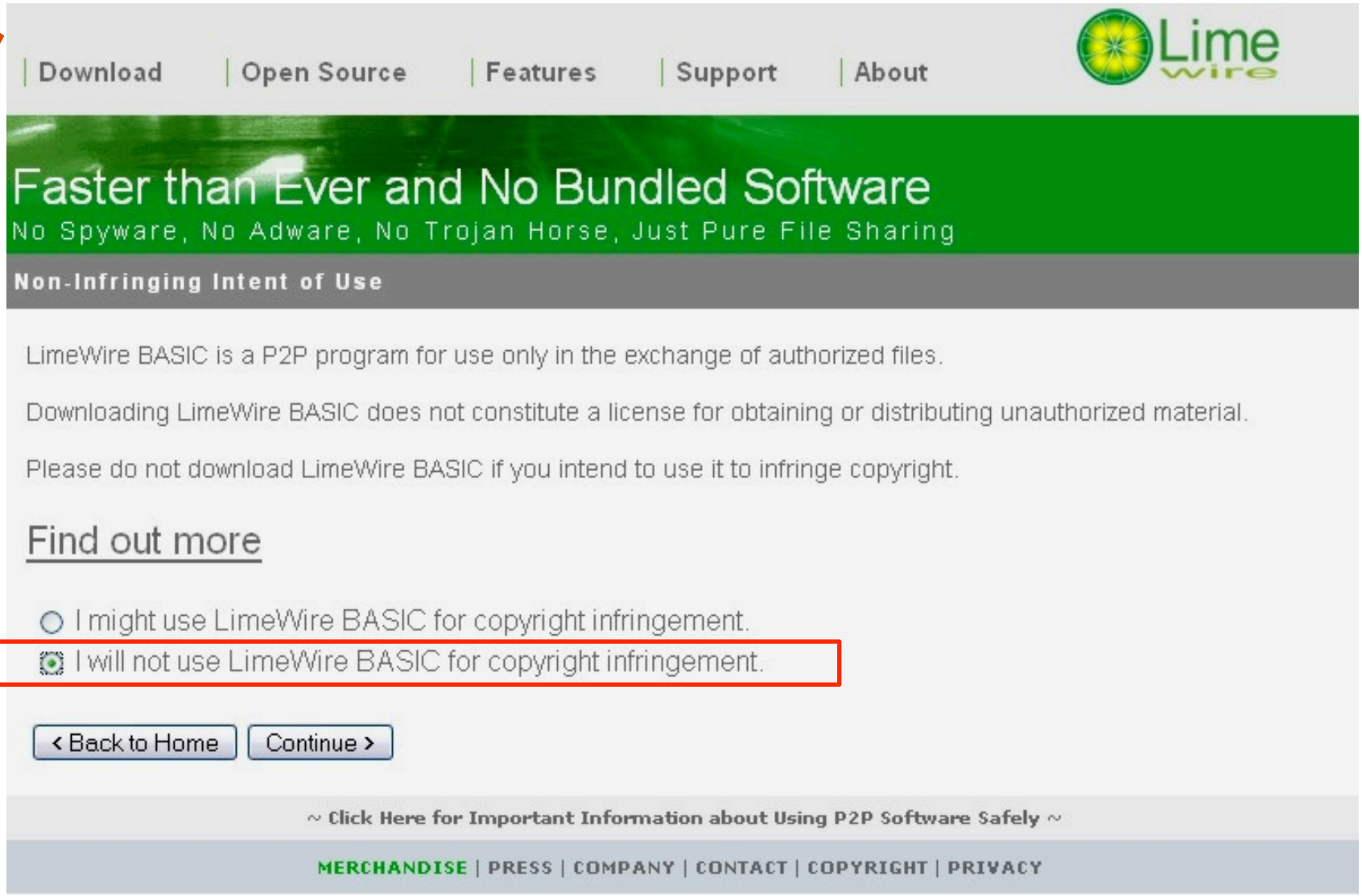
P2P networks

- Alice wants the song “foo”
  - she turns on Limewire
- Bob has the P2P application turned on and “foo” in the shared folder
- Alice’s client finds out that “foo” is on Bob’s server
- Alice initiates a direct connection with Bob and downloads “foo”

P2P

- More traffic than any other application
- Mostly media content
- Multiple issues
  - Security
  - Privacy
  - Anonimity
  - Copyright Infringement
  - Intellectual property

Copyright infringement on P2P



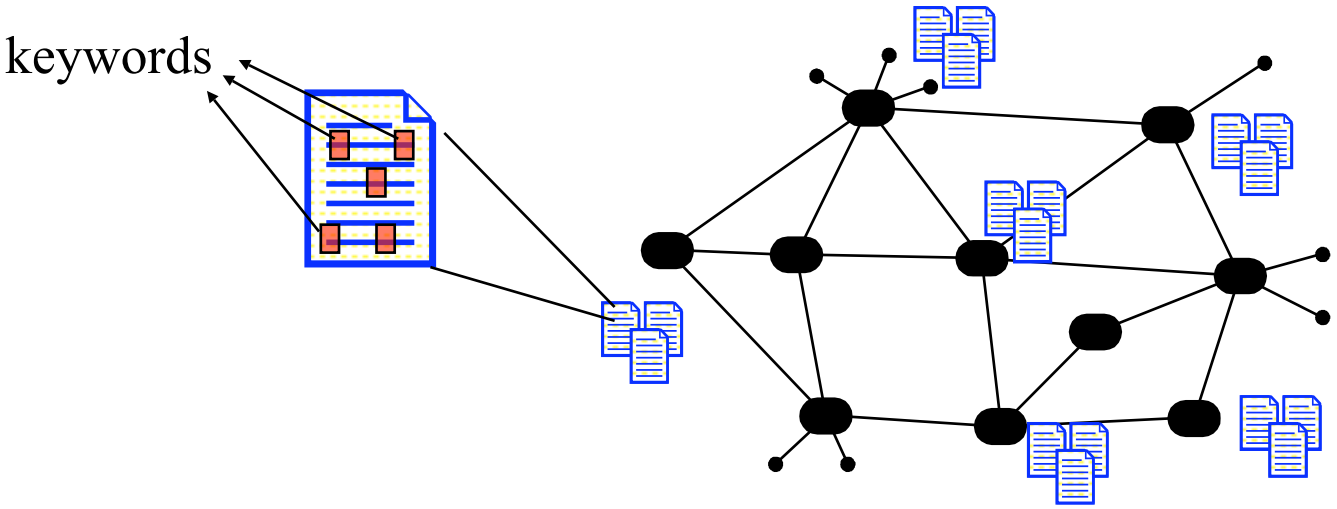
P2P

- Peer-to-Peer (P2P) systems are increasingly becoming popular.
- P2P file-sharing systems, such as Gnutella, Napster and Freenet realized a distributed infrastructure for sharing files.
- Traditionally, files were shared using the Client-Server model (e.g. http). Not scalable since they are centralized services.
- P2P uncover new advantages in simplicity of use, robustness, self organization and scalability.

P2P Information Retrieval

Problem:

"How to efficiently retrieve Information in P2P systems where each node shares a collection of documents?"



- Documents consists of keywords.
- Resembles Information Retrieval but resources are distributed now.
- Primary Data Structures such as Global Inverted Indexes can't be maintained efficiently.

P2P: Information Retrieval Issues

- Why is this more difficult than centralized IR?
  - Selection of nodes to query
    - who is up ?
  - Merging of results
  - Spam
  - Caching difficult; content changing fast

Peer-To-Peer (P2P) Search

- Distributed environment
  - Everybody does everything
- Each node in a network builds and maintains its own index
- Each node has “servent” software
  - On booting, servent pings ~4 other hosts
  - Connects to those that respond
  - Initiates, propagates and serves requests

Which hosts to connect to?

- The ones you connected to last time
- Random hosts you know of
- Request suggestions from central (or hierarchical) nameservers

- All govern system's shape and efficiency

P2P networks

- 1st generation
  - Centralized directory
  - Napster
- 2nd generation
  - Gnutella
- 3rd generation
  - FastTrack
  - Ultrapeers/SuperNodes

P2P 1st generation

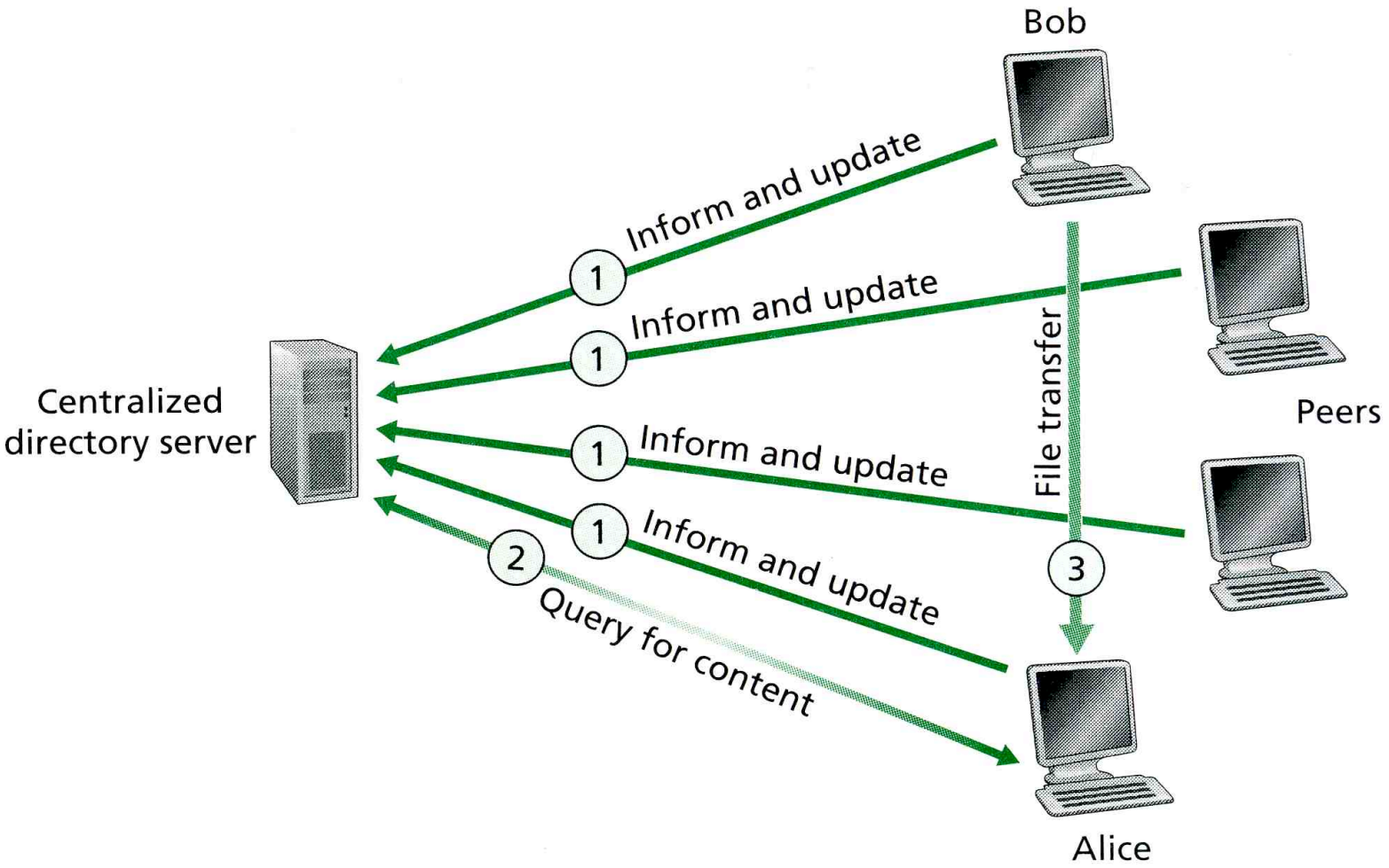


Figure 2.23 ♦ The P2P paradigm with a centralized directory

1st generation P2P

- centralized server containing most of information on the network
  - File names mapped to IPs
- single point failure
- performance bottleneck
- copyright infringement easy to track
  - Napster shut down in 2000

2nd generation P2P

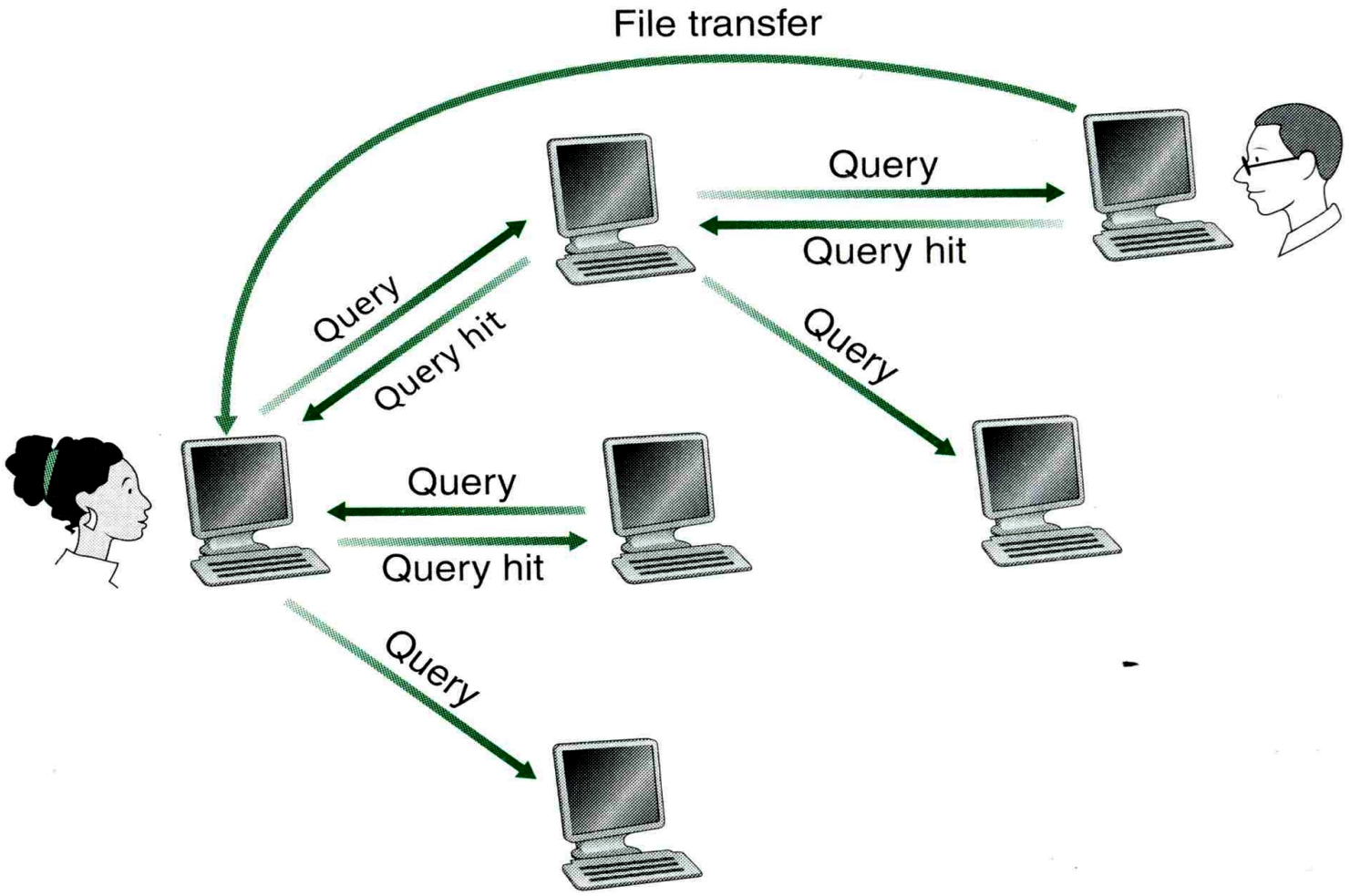


Figure 2.24 ♦ Search and file transfer in Gnutella

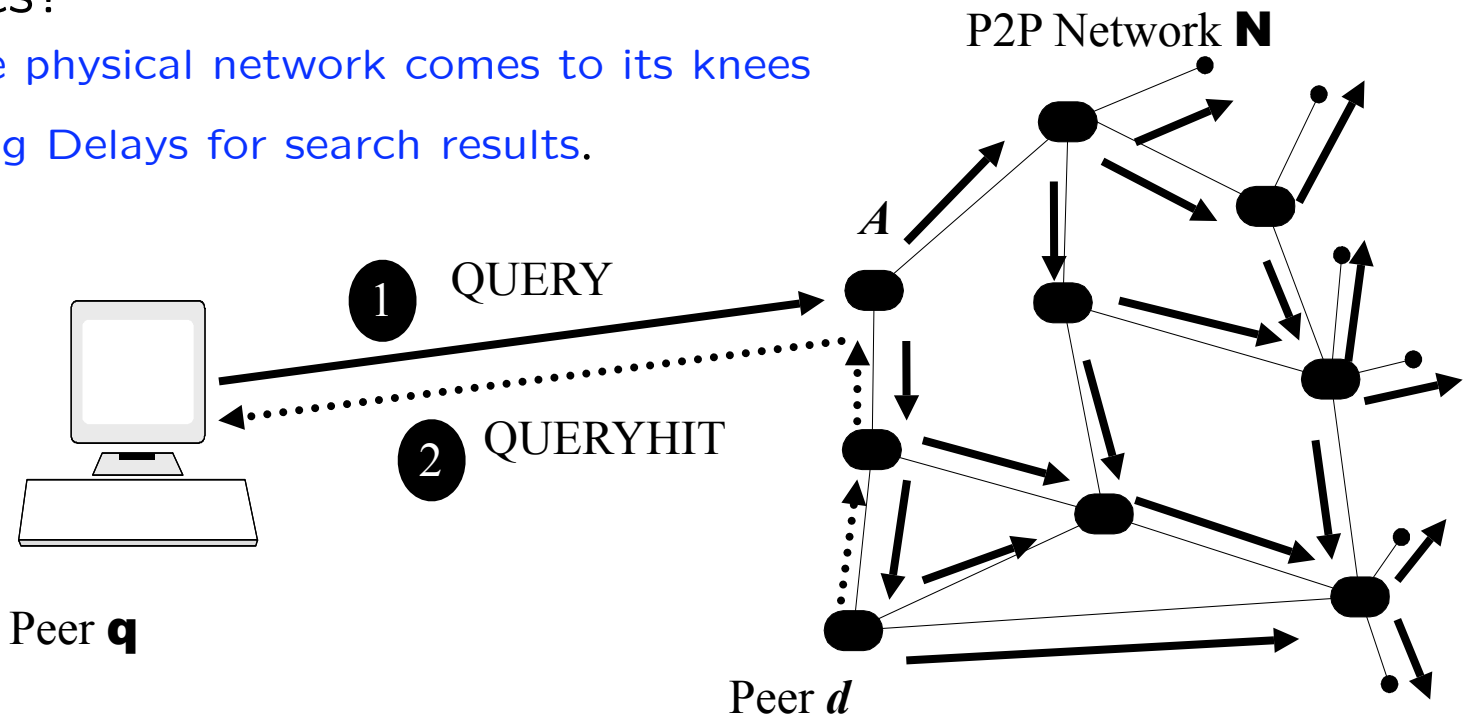
2nd generation P2P

- Gnutella
  - Limewire, Morpheus, BearShare etc
- Much harder to pursue in court
- Not-so-scalable

2nd generation P2P

1. Breadth-First Search (Gnutella)

- Query Flooding
- Each Query Message is propagated along all outgoing links of a peer using TTL (time-to-live).
- TTL is decremented on each forward until it becomes 0
- Technique for I.R in P2P systems such as Gnutella.
- Results?
  - The physical network comes to its knees
  - Long Delays for search results.



3rd generation P2P

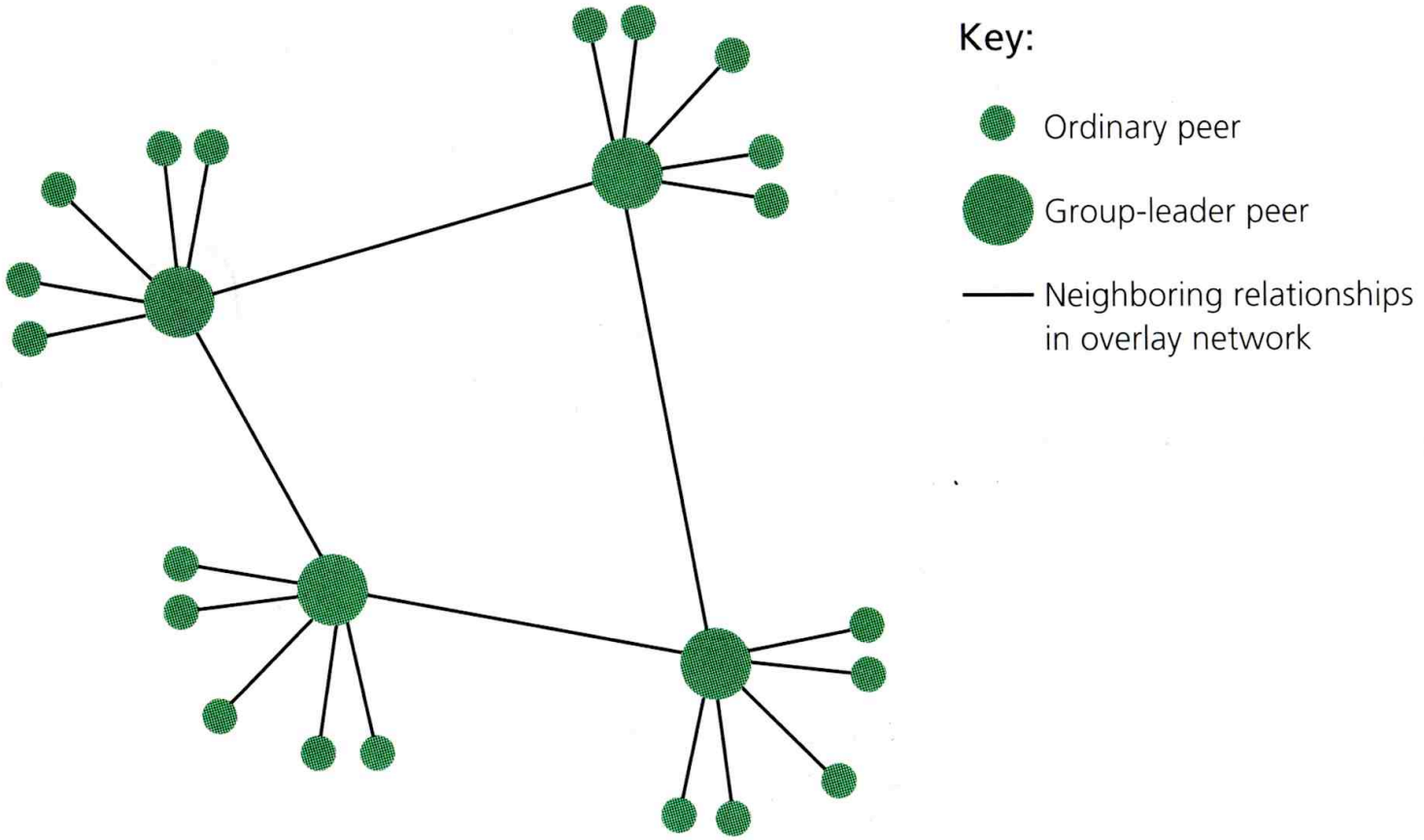


Figure 2.25 ♦ Hierarchical overlay network for P2P file sharing

ultrapeers/supernodes

- status “ultrapeer” given by
  - uptime
  - bandwidth
  - number of downloads
  - neighbors
  - need etc.
- exchange most of the info; act like Gnutella within UltraPeers
- act like a Napster for their leaves
- very scalable

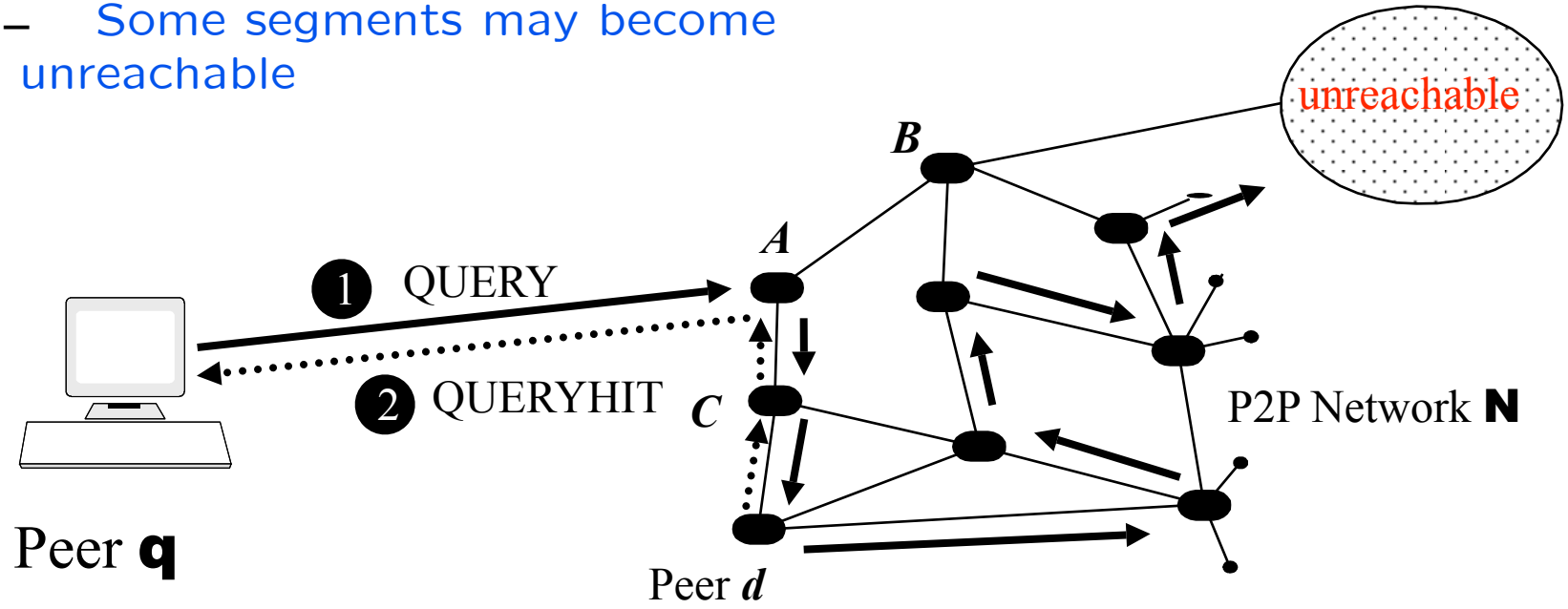
P2P - KazaA

- request queing
- incentive priorities
  - the more one uploads the better
- parallel downloading
- proprietary technology
  - encrypts all control traffic
  - numerous reverse engineering atempts
  - KazaA Lite
- hard for US organization to sue
  - patents held in Netherlands;
  - headquarters in Australia
  - developers in Estonia

Techniques for Distributed IR

Modified Random BFS

- Each Query Message is forwarded to only a fraction of outgoing links (e.g. æ of them).
- TTL is again decremented on each forward until it becomes 0.
- Results?
  - Fewer Messages but possibly less results
  - This algorithm is probabilistic.
  - Some segments may become unreachable



Techniques for Distributed I.R.

Intelligent Search Mechanism(ISM)

Idea: Each Query Message is forwarded intelligently based on what queries a peer answered in the past.

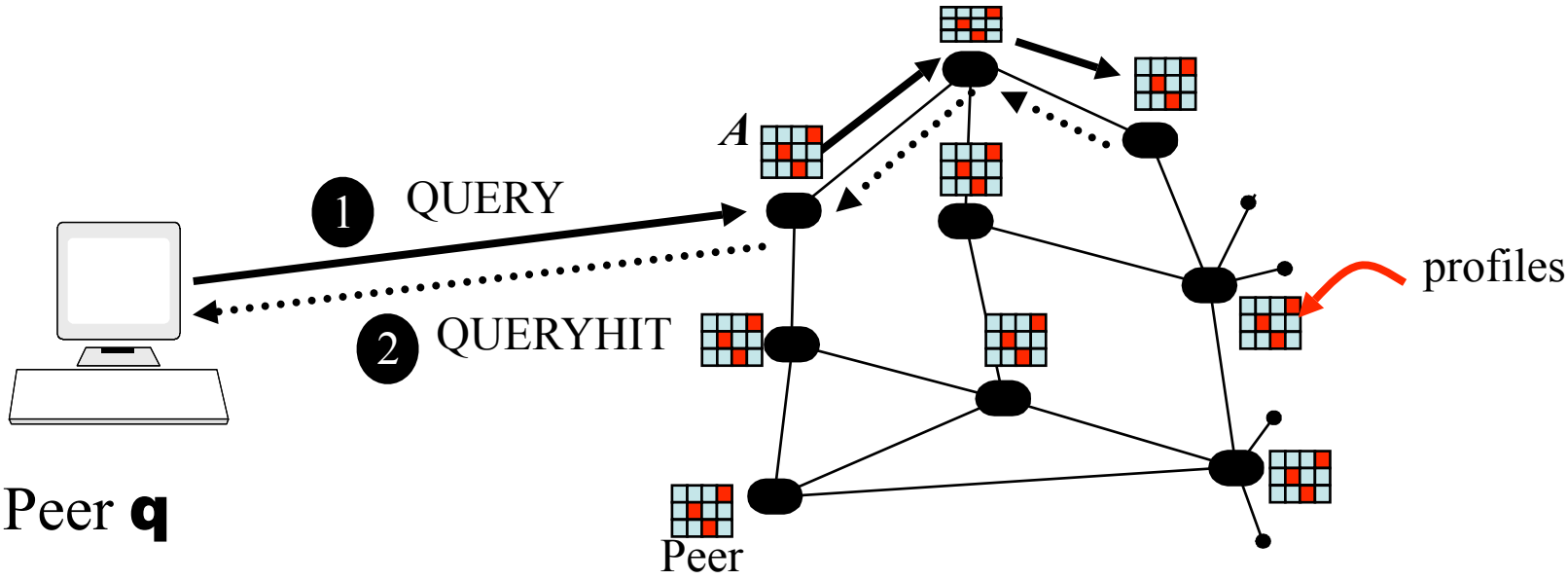
Components of ISM (for each node u)

a) Profile Mechanism, for each neighbor N(u).

b) Peer Ranking Mechanism, for ranking peers locally and send a search query only to the ones that most likely will answer.

c) Similarity Function, for finding similar search queries.

d) Search Mechanism, for propagating queries based on local indexes



Intelligent Search Mechanism (ISM)

a) Profile mechanism.

- Maintains a list of past queries routed through that host.
- Every time a QueryHit is received the table is updated

Query	GUID	Connection	timestamp
Elections Bush Clinton	G439ID	Socket1	100002222
Super Bowl San Diego	F549QL	---	100065652
***	***	***	***
Italy earthquake disaster	PN329D	Socket5	100022453

- The profile manager uses a Least Recently Used policy to keep most recent queries in repository.
- Profiles are kept for neighbors only so the cost for maintaining this cost is O(Td), T is a limiting factor per profile, d is the degree of a node

Intelligent Search Mechanism (ISM)

b) Peer Ranking Mechanism.

- Before forwarding a Query Message a peer performs an on-the-fly ranking of its peers to determine the best paths.
- We use the Aggregate Similarity of peer P_i to a query q, computed by a peer P_k as:

$$Psim_{P_k}(P_i, q) = \sum_{q_j \text{ was answered by } P_i} Qsim(q_j, q)^{\alpha}$$

Intelligent Search Mechanism (ISM)

c) Similarity Function – The cosine.

- Assume that L is a set of all words (in Profile Manager) “

e.g. L={elections, bush, clinton, super, bowl, san, diego, ... ,italy, earthquake, disaster˝

- We define an |L|-dimensional space where each query is a vector.

If q=“italy disaster” => q (vector of q) = [0,0,0,...,1,0,1]

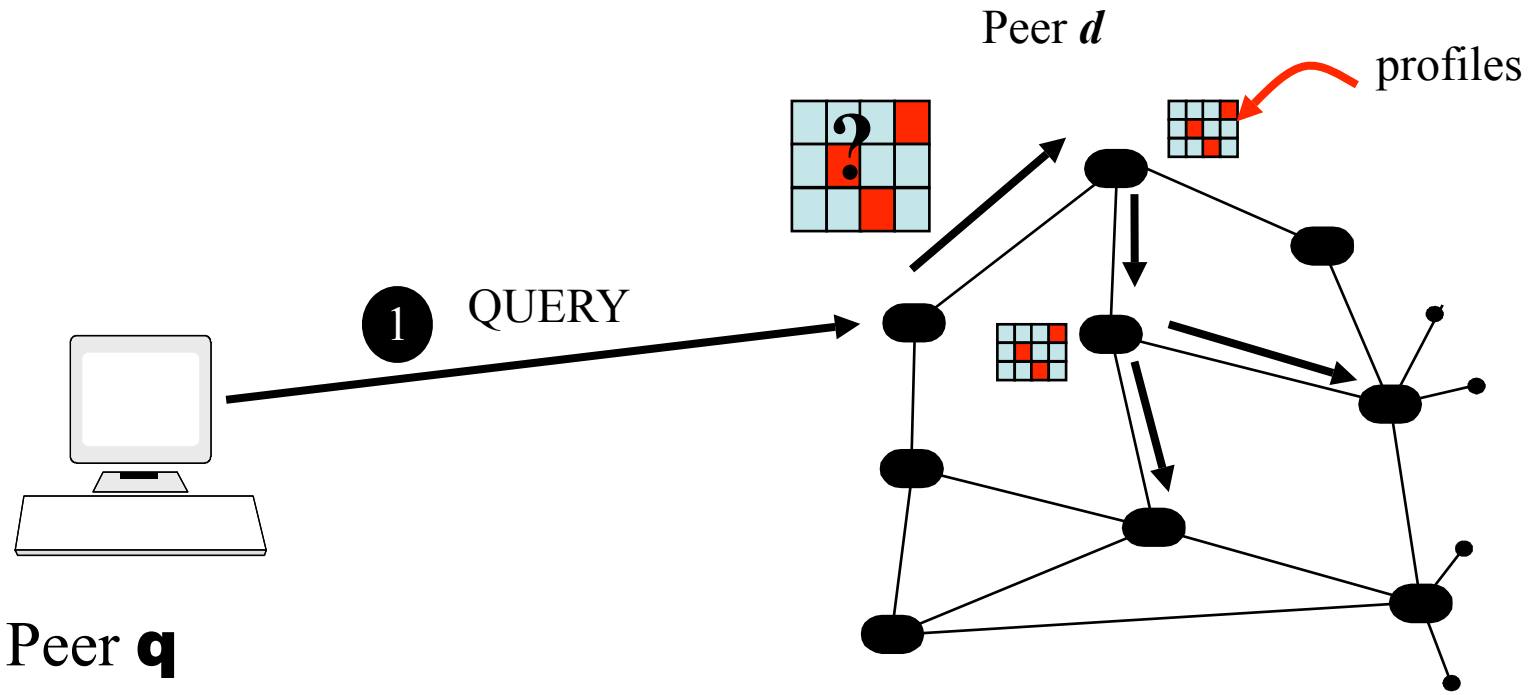
- Recall that we have a vector for each q_i stored in the Profile Manager (i.e. q_i)

$$sim(q, qi) = \cos(\vec{q}, \vec{q}i) = \frac{\vec{q}.\vec{q}i}{\| \vec{q} \|_2 * \| \vec{q}i \|_2}$$

Intelligent Search Mechanism (ISM)

Search Mechanism

- Utilizes the Peer Ranking Mechanism to forward Queries to nodes that will potentially contain the info we are looking for



Merging results

- multiple download sources
- partial dowloads, reconnecting