

Search, Pollution, and Poisoning in P2P File-Sharing Networks

John Chuang

School of Information Management and Systems

University of California at Berkeley

chuang@sims.berkeley.edu

<http://p2pecon.berkeley.edu/>

Guest Lecture for IS290: Search Engines, October 3 2005

P2P Search and WWW Search

- Similarities

- Scale and scope: Kazaa alone has 4 petabytes of data shared by 3 million peers (2004)

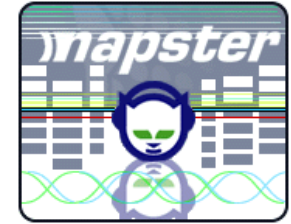
- Differences

- Highly dynamic: peers come and go
 - Short session durations → high content volatility

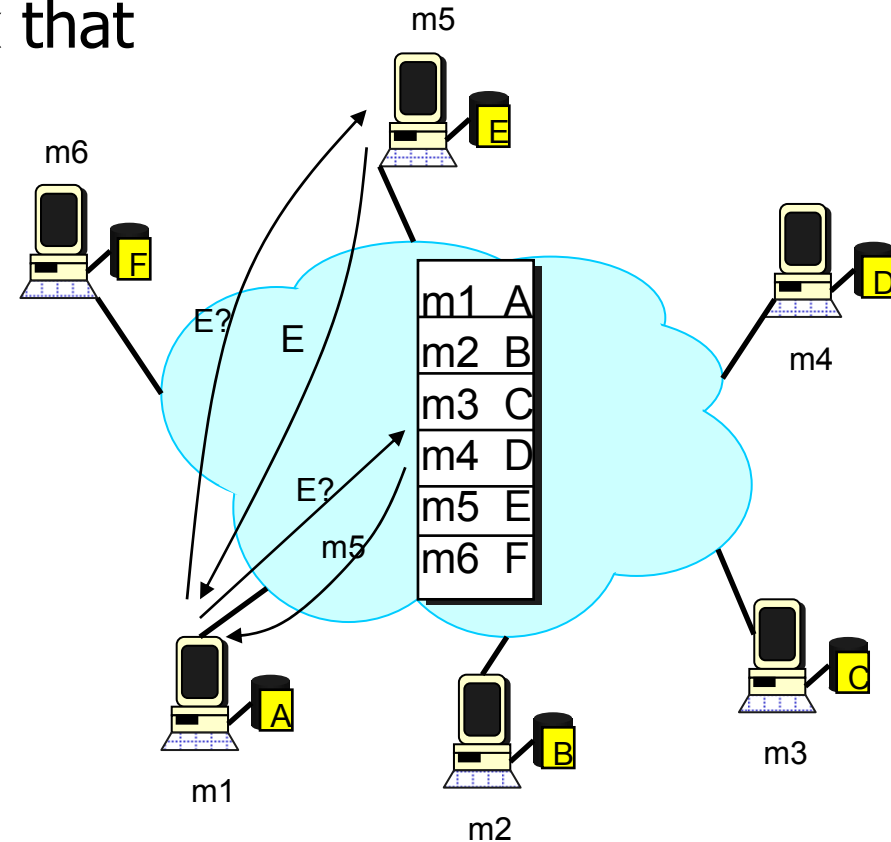
P2P File-Sharing Networks

- 1st generation: centralized index
 - e.g., Napster
- 2nd generation: decentralized indices
 - e.g., Gnutella v0.4, Freenet
- 3rd generation: hierarchical
 - e.g., FastTrack (KaZaA, Grokster, Morpheus), eDonkey2000, Gnutella v0.6
- 4th generation?: structured topologies
 - e.g., Overnet using Kademlia DHT
- Note: BitTorrent has no built-in search mechanism; various darknet proposals for small-scale “F2F” networks

Napster



- Maintains a centralized index that maps files to machines
- How to find a file
 - Query the index system → return a list of peers that store the requested file
 - Transfer the file directly from peer(s)
- Advantage:
 - Simplicity: easy to implement sophisticated search engines on top of the index system
- Disadvantage:
 - Single point of failure

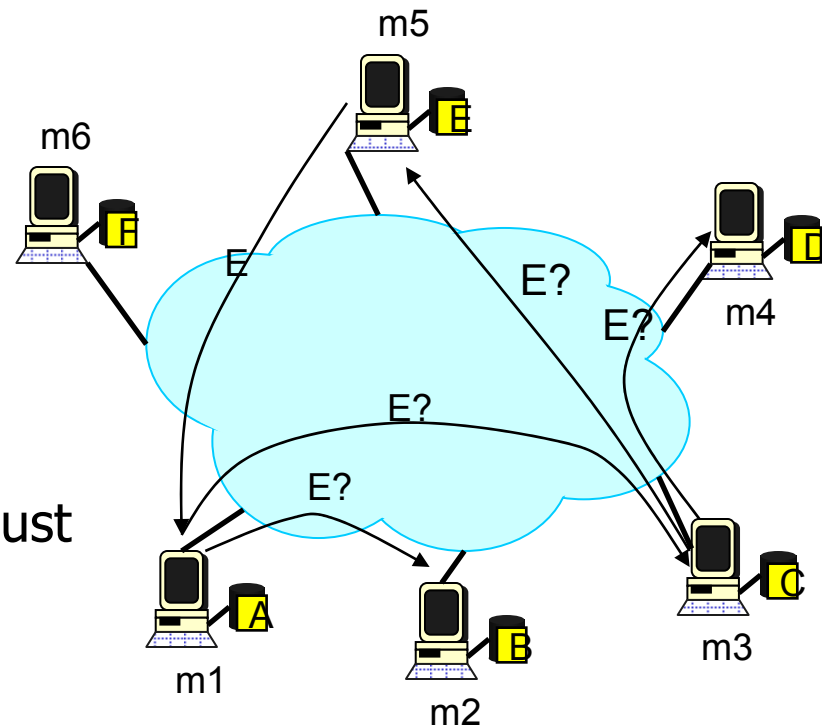


Gnutella (v0.4)



gnutella.com

- Flood the request
- How to find a file:
 - Send request to all neighbors
 - Neighbors recursively propagate the request
 - Eventually a machine that has the file receives the request, and it sends back the answer
- Advantages:
 - Totally decentralized, highly robust
- Disadvantages:
 - The entire network can be swamped with a request
 - Can be alleviated using TTLs, but can then fail to locate files (and still high resource usage)

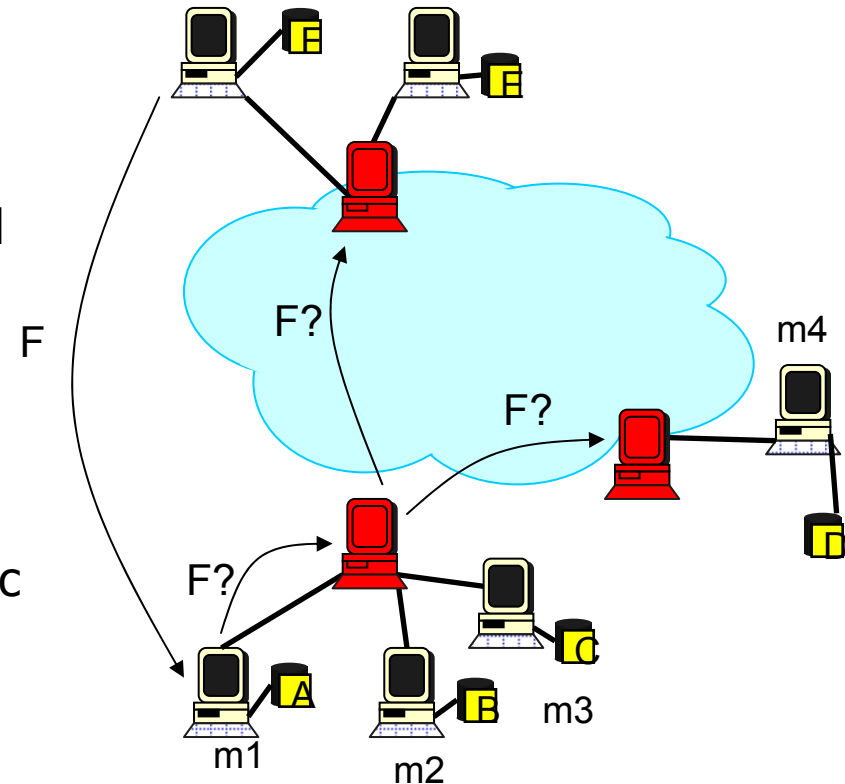


Assume: m1's neighbors are m2 and m3;
m3's neighbors are m4 and m5;...

Hierarchical networks



- Use two-level hierarchy
 - Some nodes are elected as “super nodes” or “ultra-peers”
 - Each ultra-peer serves as centralized index for a portion of the network
 - If an ultra-peer does not know where to find an item, query is forwarded to other ultra-peers
- Advantage:
 - Reduce the amount of network traffic compared to “naïve” flooding
- Disadvantage:
 - Ultra-peers vulnerable to attacks
 - Potential convergence problems when ultra-peers leave abruptly
- Used in FastTrack (KaZaA, Grokster, Morpheus), eDonkey2000, Gnutella v0.6



Assume red nodes are ultra-peers

A different perspective

- Copyright owners might prefer to minimize the effectiveness of search in P2P file-sharing networks
 - Viable technological alternative to legal recourse?
- A number of companies specialize in injection of noise into P2P file-sharing networks on behalf of copyright owners
 - e.g., Overpeer, Retspan, Macrovision, ...
- Our question: what is the effect of pollution and poisoning on the availability of content in P2P file-sharing networks?
 - Christin, Weigend, and Chuang, ACM EC 2005

Pollution vs. Poisoning

- Network pollution

- *Accidental* injection of unusable or low quality files
 - Happens with most (all?) content
 - Truncated, poorly encoded, ...
 - Difficulties in properly “ripping” content

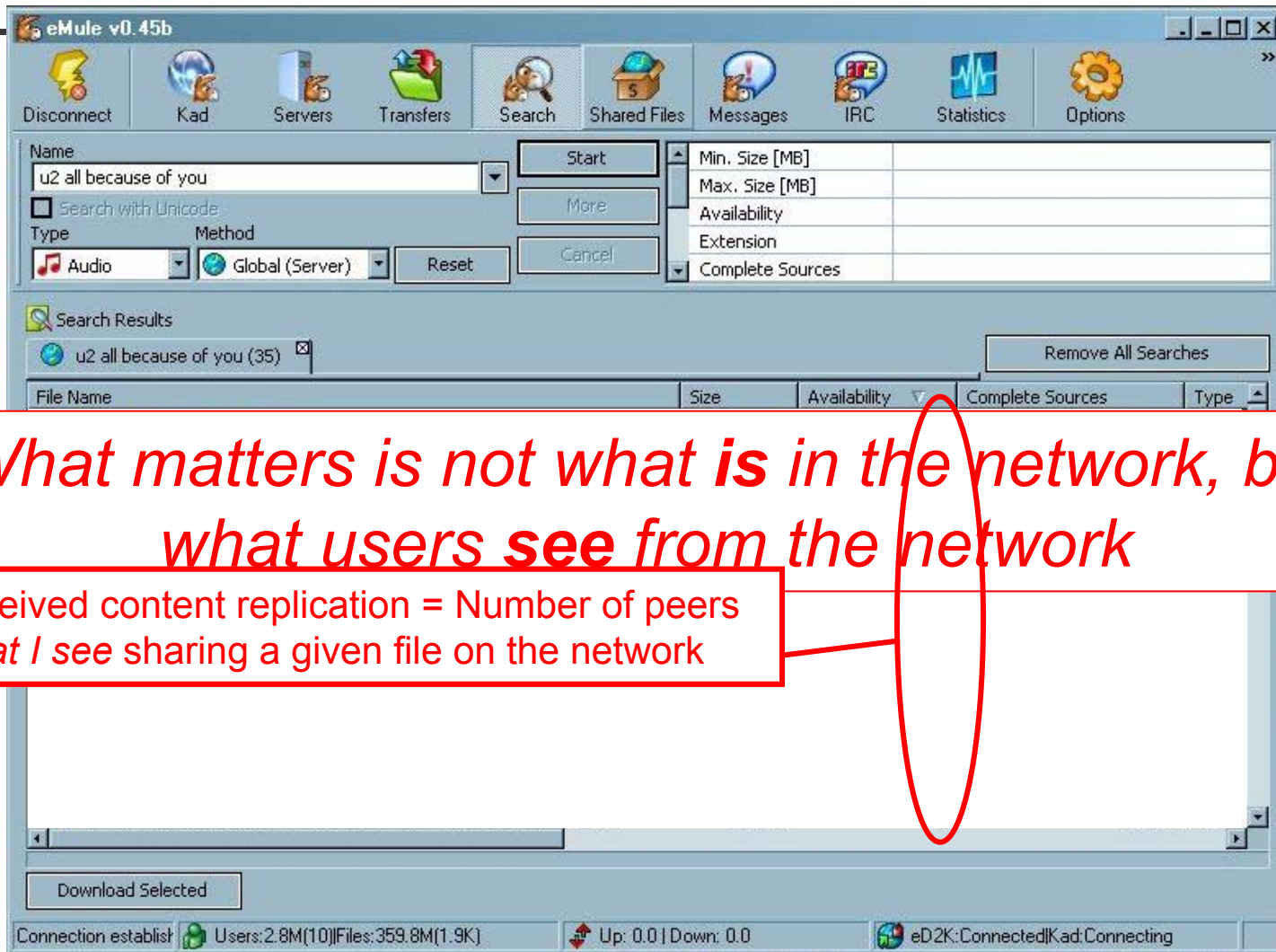
- Item poisoning

- *Deliberate* injection of decoys to render usable files hard to find
 - Targets specific content
 - e.g., “American Life” by Madonna

Questions

- Above which level does pollution pose serious problems?
- Which (if any) poisoning techniques are effective?
 - Flooding?
 - More elaborate techniques?
- We'll look at the most popular P2P networks
 - FastTrack (KaZaA), eDonkey, Overnet, Gnutella
 - not BitTorrent – does not have built-in search mechanism

Availability vs. perceived availability



Differing perceptions of content

- Ideally all P2P nodes should have same view of content available on the network
- In practice, different nodes have very different perceptions of content availability
 - Peers coming and going → Content volatility
 - Size of the network/decentralized nature imposes fish-eye view
- User view of the network conditioned by query returns
- Query returns highly dependent on P2P network topology

P2P topologies

- Most modern P2P networks use 2-level hierarchical structure
 - Leaf nodes
 - Hubs (a.k.a. supernodes, ultrapeers, servers)
 - Higher processing power, link capacity, longer uptime...
 - Act as a centralized index for a number of leaf nodes
- Exception: Overnet
 - Distributed Hash Table (all peers are equal)
 - However, Overnet clients are also part of the eDonkey network

Differences in topological structures

	eDonkey	FastTrack	Gnutella
# of hubs	40—90	25,000—	10,000— 100,000
# of nodes	≈ 2,800,000	≈ 2,500,000	≈ 1,000,000
Fraction of hubs	≈ 0.00002	≈ 0.015	≈ 0.05
Avg. leaf-hub connection lifetime	≈ 24 hours	≈ 30 minutes	≈ 90 minutes
Leaf promotion	Voluntary	Election	Election

Semi-centralized network

Hubs are much more stable

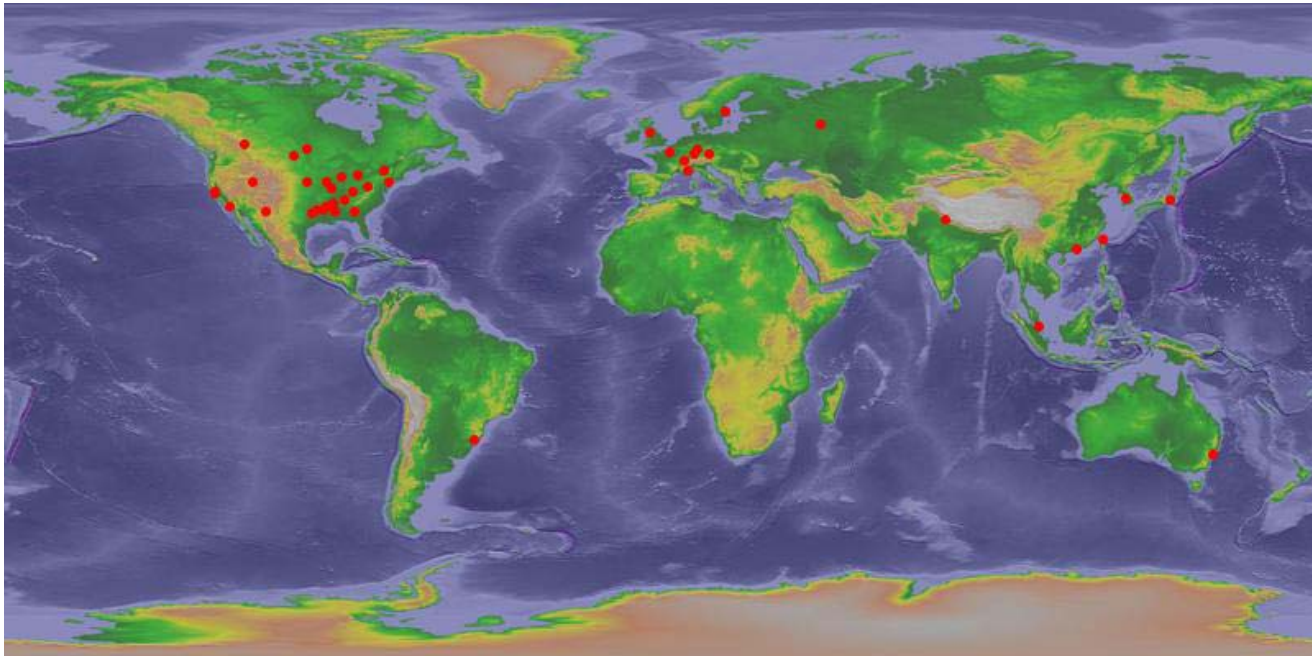
Methodology

Perception of availability depends on time and origin of a query

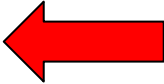
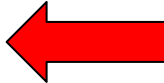
- Need to measure from different vantage points and at different times
1. Measure content availability *in absence* of poisoning
 2. Evaluate effect of pollution and poisoning on measured data by numeric simulation

Measurement infrastructure

- giFT-FastTrack and MLDonkey clients
 - Linux console (text-based) applications
 - Allows for scripting
- Easy to run large scale experiments
 - 50 host machines over 18 different countries (PlanetLab)



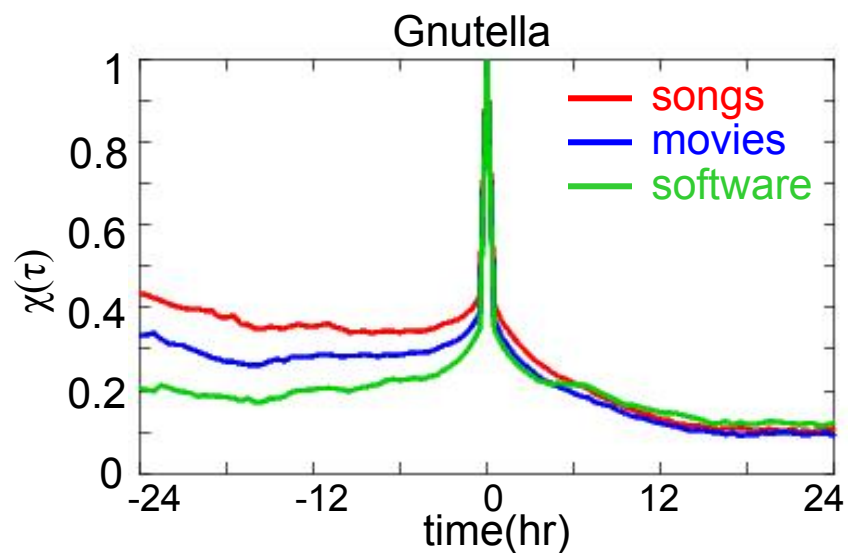
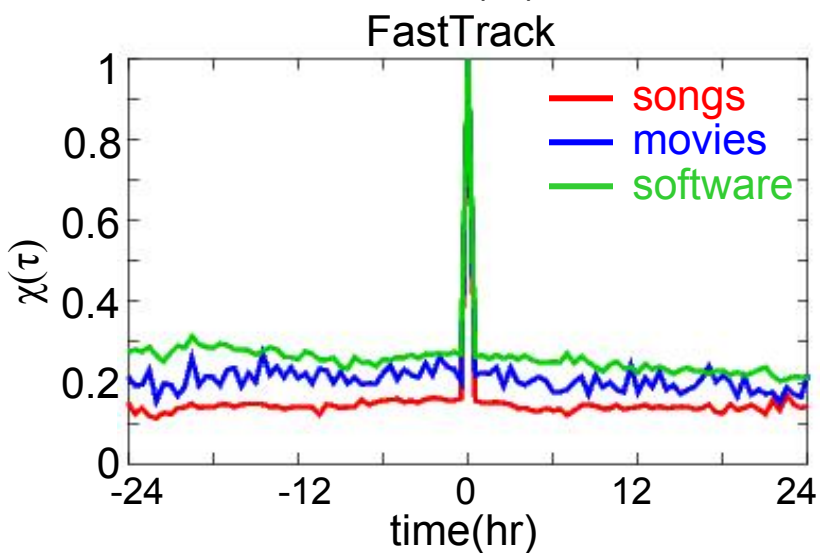
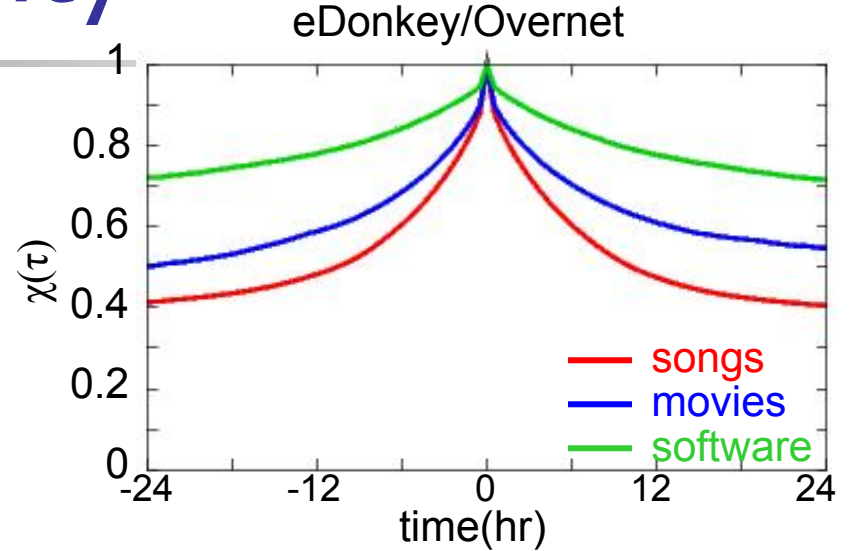
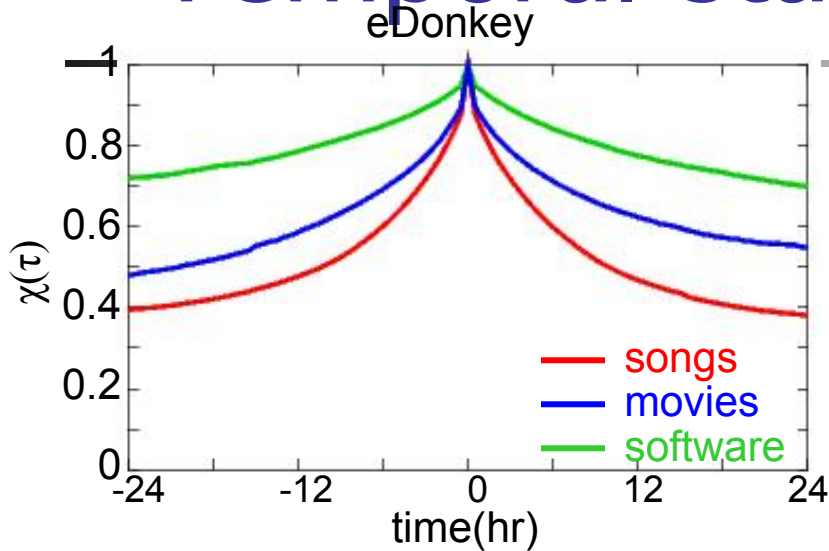
Measured data

- Network metrics likely to have an impact on users' decisions to use or abandon a given P2P network
 - Number of responses to a query
 - Query response time
 - Content stability
 - Temporal stability 
 - Spatial stability
 - Content replication 
 - Download completion time

Temporal stability

- Assess how the users' perception of the available content changes over time
- $\chi(\tau)$: average probability (over all times, all clients) that an item (specific file) returned at a given time T is also returned at time $T+\tau$
 - “what is the probability that I will see tomorrow a file that I see now?”
 - “what is the probability that a file that I see in the network now was already available to me three hours ago?”
- Could be very heavily impacted by poisoning attacks

Temporal stability



Perceived content replication

The screenshot shows the eMule v0.45b interface. The search results table is as follows:

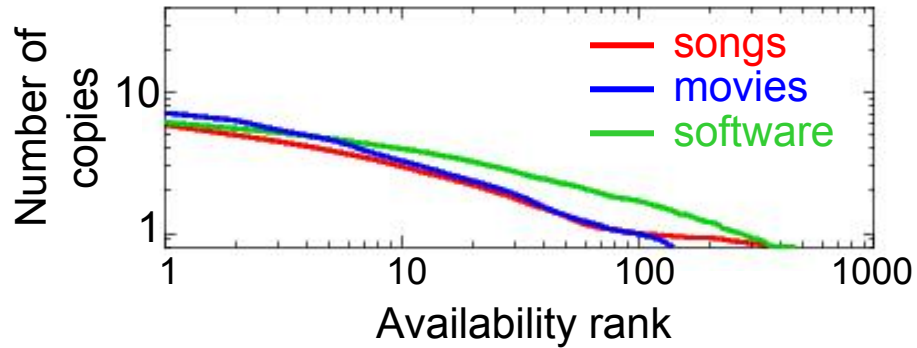
File Name	Size	Availability	Complete Sources	Type
U2 - All Because Of You.mp3	6.31 MB	32	93%	Audio
06.- U2 - all because of you - [EMG].mp3	6.31 MB	20	90%	Audio
U2 - How To Dismantle An Atomic Bomb - 06 -All Because Of You.mp3	6.31 MB	11	54%	Audio
U2 - All because of you.mp3	6.36 MB	7	85%	Audio
U2 - All Because of You.mp3	6.31 MB	5	80%	Audio
U2 - How To Dismantle An Atomic Bomb - 06 - All Because Of You.mp3	6.31 MB	2	100%	Audio
06 - U2 - all because of you - [EMG].mp3	6.32 MB	2	100%	Audio
U2 - All Because of You.mp3	6.31 MB	2	100%	Audio
U2 - All Because of You.mp3	6.31 MB	1	100%	Audio
U2 - All Because of You.mp3	6.31 MB	1	100%	Audio
U2 - All Because of You.mp3	6.31 MB	1	100%	Audio
U2 - All Because of You.mp3	6.31 MB	1	100%	Audio
06.- U2 - all because of you - [EMG].mp3	6.32 MB	1	100%	Audio
U2 - All Because of You.mp3	3.35 MB	1	0%	Audio
U2 - How To Dismantle An Atomic Bomb - 06 - All Because of You.mp3	6.31 MB	1	100%	Audio
U2 - How To Dismantle An Atomic Bomb - 06 - All Because of You.mp3	6.31 MB	1	100%	Audio

A red oval highlights the 'Availability' column, and a red box with a pointer contains the following text:

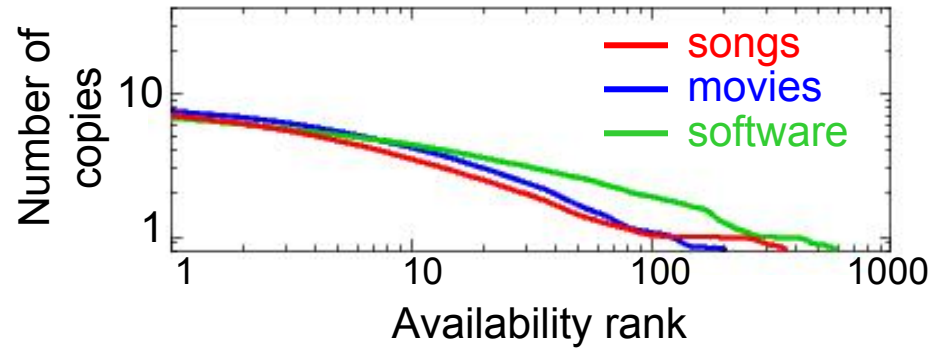
Perceived content replication = Number of peers that I see sharing a given file on the network

Perceived content replication

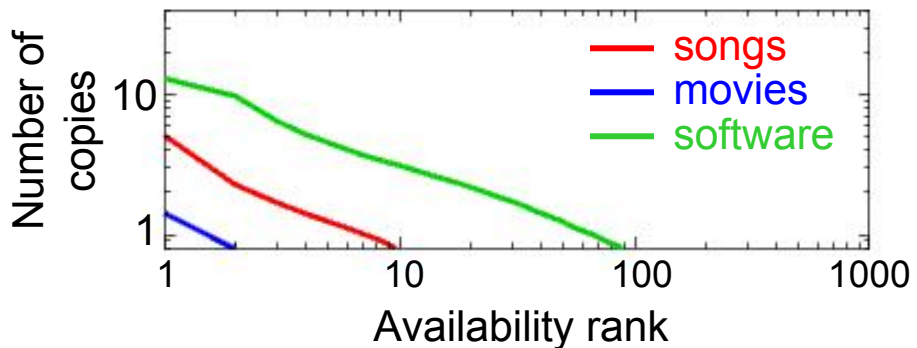
eDonkey



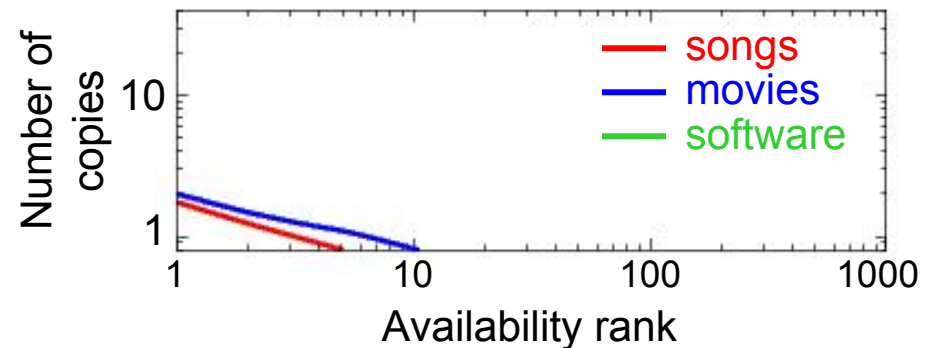
eDonkey/Overnet



FastTrack



Gnutella



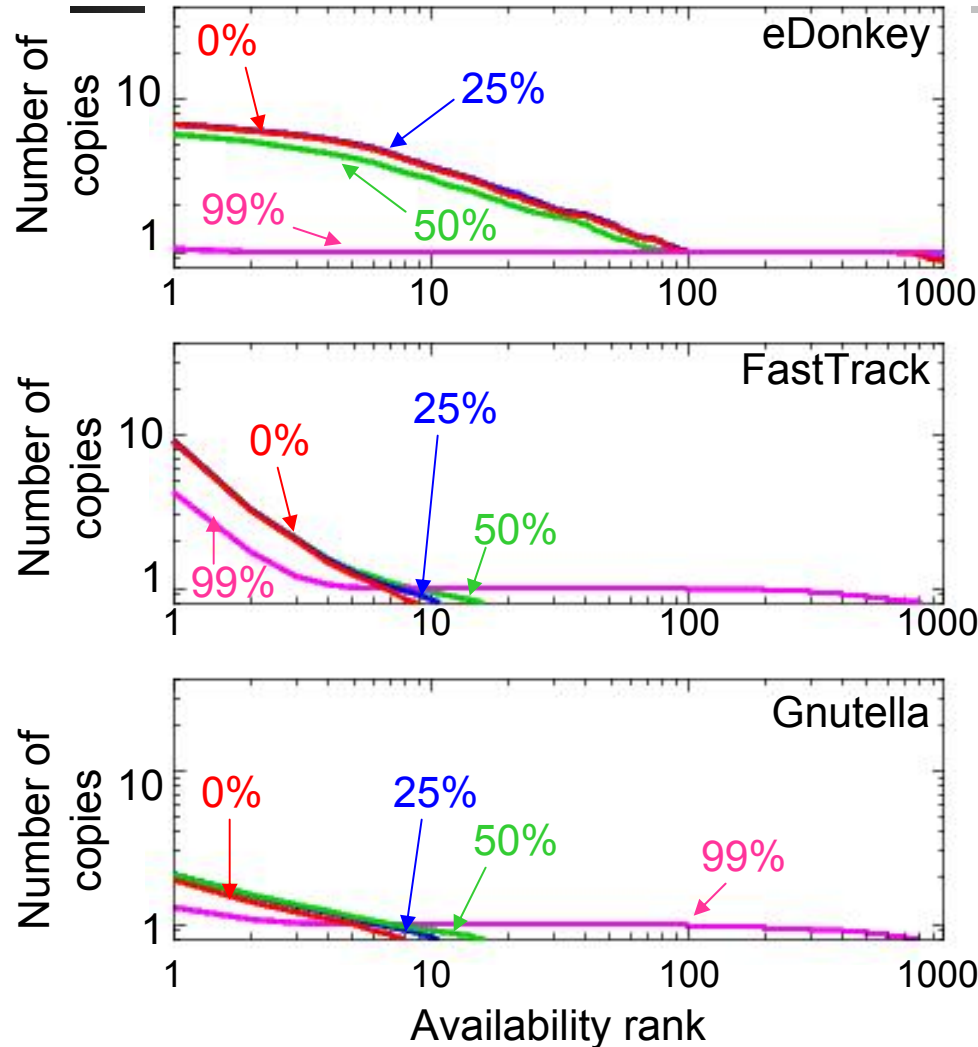
Summary of measurements in the absence of poisoning

- Semi-centralized topologies (eDonkey)
 - Content remains present in the network for a while
 - Faster responses to queries
- FastTrack and Gnutella
 - Relatively low content stability
 - content comes and goes frequently
 - Apparently high levels of pollution
 - even when no poisoning
 - Manage to only download a few files
 - Confirms findings of (Liang *et al.*, 2005)

Effects of pollution

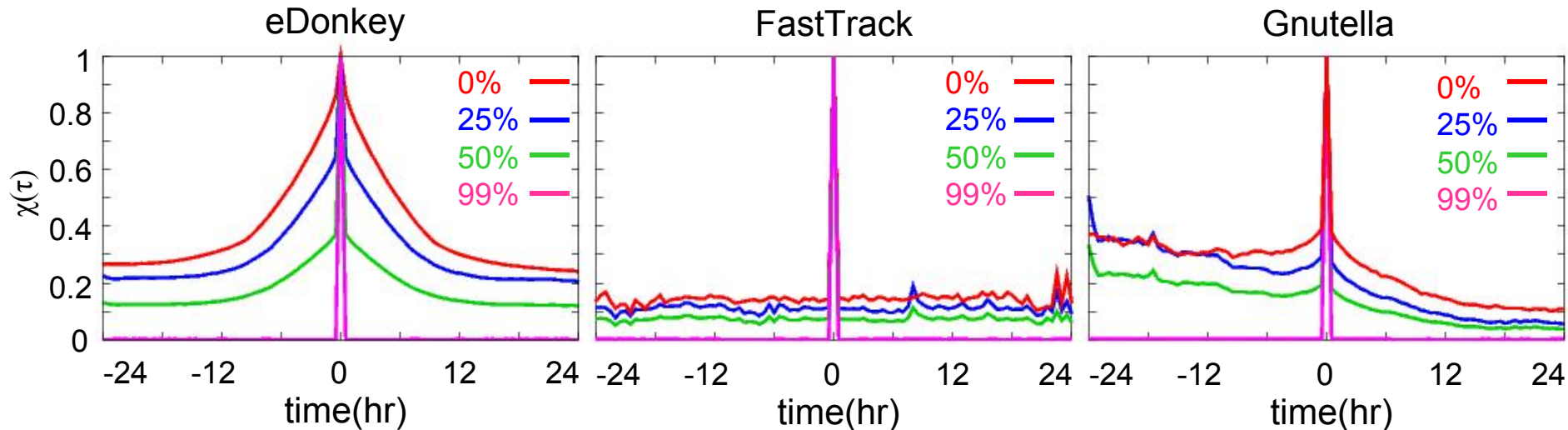
- Pollution modeled as injection of random noise in the system
 - Make $x\%$ of the query returns (uniformly) random for each measurement sample
 - Neglects propagation effects of polluted content
- Simplest poisoning technique (flooding) is nothing more than pollution at high levels
 - Should not, *in theory*, reduce availability of useful files
 - In practice, number of query returns is limited
 - FastTrack example:
 - At most 200 returns for a given query
 - No more than 5 queries in a row

Pollution and perceived availability



- Pollution only harmful at (very) high levels
- However, decoys *may* drive usable files out of the query returns
- Poisoning by flooding not particularly efficient
 - e.g., need to insert 99 times as many decoys as existing files
 - ... at each hub

Flooding signature



- High-levels of pollution (or poisoning by flooding) completely destroys temporal stability
- Easy to thwart by giving precedence to items that have been seen in the network for some time

Alternatives to flooding

- More advanced poisoning techniques can be much less expensive and more efficient than flooding
 - A (rather detailed) list of attacks is available in a patent application from Macrovision
 - Discussed at <http://mvsn-patent-app.notlong.com>
 - Chunk corruption
 - Malicious routing
 - Skewing perceived availability to bias users towards downloading useless content
 - ...

Targeting perceived availability

eMule v0.45b

Disconnect Kad Servers Transfers Search Shared Files Messages IRC Statistics Options

Name: u2 all because of you Start Min. Size [MB] Max. Size [MB] Availability Extension Complete Sources

Search with Unicode More Cancel

Type: Audio Method: Global (Server) Reset

Search Results: u2 all because of you (35) Remove All Searches

File Name	Size	Availability	Complete Sources	Type
U2 - All Because of You - Poisoned edition (rare!).mp3	6.31 MB	98	100%	Audio
U2 - All Because of You - Poisoned edition (rare!).mp3	6.31 MB	98	100%	Audio
U2 - All Because of You - Poisoned edition (rare!).mp3	6.31 MB	98	100%	Audio
U2 - All Because of You - Poisoned edition (rare!).mp3	6.31 MB	98	100%	Audio
U2 - All Because of You - Poisoned edition (rare!).mp3	6.31 MB	98	100%	Audio
U2 - All Because of You - Poisoned edition (rare!).mp3	6.31 MB	98	100%	Audio
U2 - All Because of You - Poisoned edition (rare!).mp3	6.31 MB	98	100%	Audio
U2 - All Because of You - Poisoned edition (rare!).mp3	6.31 MB	98	100%	Audio
U2 - All Because of You - Poisoned edition (rare!).mp3	6.31 MB	98	100%	Audio
U2 - All Because of You - Poisoned edition (rare!).mp3	6.31 MB	98	100%	Audio
U2 - All Because of You - Poisoned edition (rare!).mp3	6.31 MB	98	100%	Audio
U2 - All Because of You - Poisoned edition (rare!).mp3	6.31 MB	98	100%	Audio
U2 - All Because of You - Poisoned edition (rare!).mp3	6.31 MB	98	100%	Audio
U2 - All Because of You - Poisoned edition (rare!).mp3	6.31 MB	98	100%	Audio
U2 - All Because of You - Poisoned edition (rare!).mp3	6.31 MB	98	100%	Audio
U2 - All Because of You - Poisoned edition (rare!).mp3	6.31 MB	98	100%	Audio
U2 - All Because of You - Poisoned edition (rare!).mp3	6.31 MB	98	100%	Audio
U2 - All Because of You - Poisoned edition (rare!).mp3	6.31 MB	98	100%	Audio
U2 - All Because of You - Poisoned edition (rare!).mp3	6.31 MB	98	100%	Audio
U2 - All Because of You - Poisoned edition (rare!).mp3	6.31 MB	98	100%	Audio
U2 - All Because of You - Poisoned edition (rare!).mp3	6.31 MB	98	100%	Audio
U2 - All Because of You - Poisoned edition (rare!).mp3	6.31 MB	98	100%	Audio
U2 - All Because of You - Poisoned edition (rare!).mp3	6.31 MB	98	100%	Audio
U2 - All Because of You - Poisoned edition (rare!).mp3	6.31 MB	98	100%	Audio

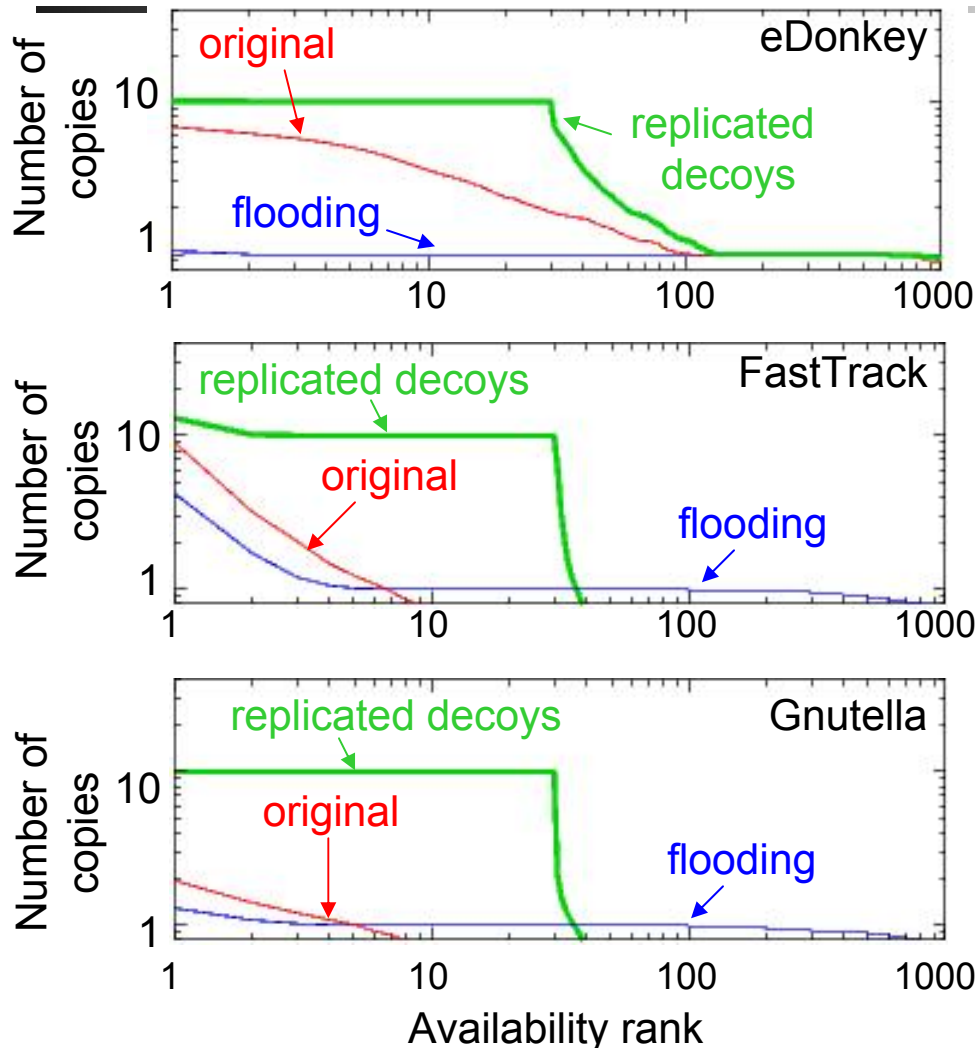
Download Selected

Connection established Users: 2.8M(10) | Files: 359.8M(1.9K) Up: 0.0 | Down: 0.0 eD2K: Connected | Kad: Connecting

Targeting perceived availability

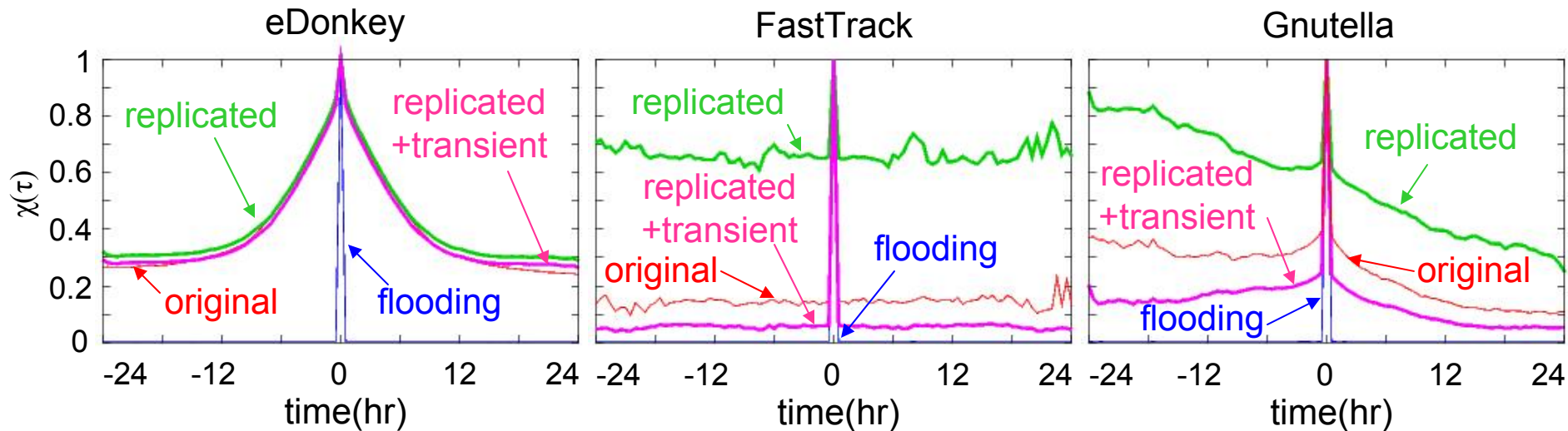
- Inject a few highly replicated decoys rather than random files
- Can in addition make replicated decoys harder to detect by frequently changing them (transient decoys)

Replicated decoy injection



- Insert 30 decoys with the same number of copies as most replicated file
- Drives useful files out of the picture
- Here only requires about 300 decoys
 - as opposed to ~9900 for flooding

Temporal signatures



- Using permanent replicated decoys leaves a rather obvious signature on the temporal stability
- Can be solved by frequently changing the (replicated) decoys

Poisoning antidotes

- Ranking by availability
 - Simplest technique
 - Efficient against random noise (if no propagation)
- Static reputation system
 - “File X is useless,” “IP address Y injects useless content”
 - Needs manual input, far from comprehensive
 - <http://www.jugle.net>, <http://bitzi.com>
- Dynamic ((semi-)automated) reputation system
 - Weighs reputation of a file as a number of factors
 - Manual input
 - Time present in the system
 - Semi-automate ban of poisoning sources
 - Unlikely such systems are *currently* deployed

Antidotes and their effectiveness

	Pollution	Flooding	Replicated decoys	Replicated, transient decoys
Ranking by number of replicas found	Yes	Somewhat	No	No
Static reputation	Somewhat	No	Yes	No
Dynamic reputation	Somewhat	Somewhat	Yes	Somewhat

The poisoning arms race

P2P designers

- Need to use several antidotes in conjunction
 - e.g., ranking by number of replicas with reputation
- Efficiency of reputation systems improved by looking at statistical characteristics
 - Temporal stability signatures

Copyright holders

- Brute force never a bad choice
 - Can be devastating if used with proper (combination of) strategies
- Clever techniques can use the reputation system to catalyze poisoning
 - False positives
 - False negatives

Summary

- Network topology plays a crucial role in how users perceive content
 - (Semi-)centralized topologies provide more stable content
- Easy to combat (involuntary) pollution
 - E.g., ranking results by number of replica found
- More advanced poisoning strategies harder to thwart
 - Arms race between poisoning techniques and reputation systems

Questions?

N. Christin, A. Weigend, and J. Chuang,
“Content Availability, Pollution and Poisoning
in Peer-to-Peer File Sharing Networks.” *Proc.
ACM E-Commerce Conference (EC'05)*.
Vancouver, BC, Canada. June 2005.

Paper available at <http://p2pecon.berkeley.edu>