Content-Based Search in Internet-Scale Peer-to-Peer Systems

by

Demetris Zeinalipour

Visiting Lecturer
Department of Computer Science
University of Cyprus

Thursday, December 28th, 2006
Royal Institute of Technology (KTH)
Stockholm, Sweden

http://www.cs.ucy.ac.cy/~dzeina/

Presentation Goals

- To provide an **overview** of **Content-Based Search Algorithms** in P2P systems with an emphasis on **ISM**.
- To provide an **overview** of **Topologically-Aware overlay construction mechanisms** in P2P systems with an emphasis on **DDNO**.
- To present other **research activities** that our group is currently involved in.

References Related to this Talk

- **"pFusion: An Architecture for Internet-Scale Content-Based Search and Retrieval" by**

  D. Zeinalipour-Yazti, V. Kalogeraki, D. Gunopulos, IEEE Transactions on Parallel and Distributed Systems, (**IEEE TPDS**), accepted, 2006.

- **"Structuring Topologically-Aware Overlay Networks using Domain Names",**

  D. Zeinalipour-Yazti, V. Kalogeraki, **Computer Networks** (Comnet), Elsevier Publications, Volume 50, Issue 16 , pp. 3064-3082, 2006.

- **"Exploiting Locality for Scalable Information Retrieval in Peer-to-Peer Systems",** D. Zeinalipour-Yazti, V. Kalogeraki and D. Gunopulos, Information Systems (InfoSys), Volume 30, Issue 4, Pages 277-298, 2005.

Introduction to Peer-to-Peer

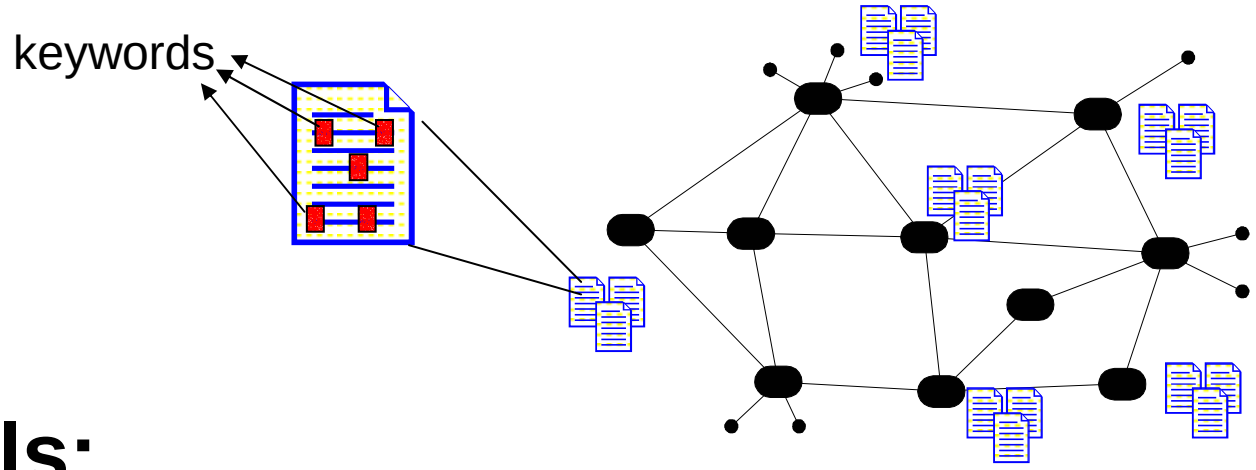
- The P2P Computing paradigm became a powerful model for developing **infrastructure-less Internet-Scale** systems.
- **Internet-Scale:** Large number of geographically spread nodes.
- Fascinating Applications :
  - **File-sharing** (e.g. Napster, Gnutella, eDonkey,...)
  - **Internet Telephony** (e.g. Skype from Kazaa team)
  - **Spam Detection Networks** (e.g. SpamNet)
  - **Web Caching** (e.g. SQUIRREL based on Pastry)
  - **Web Anonymizers** (e.g. Tarzan)
  - **Distributed Computing** (e.g. Seti@Home)
  - **P2P Online Gaming (e.g. Sony Playstation), Content Distribution Networks (e.g. Coral)**

A) Content-Based Search in P2P Systems

Problem Definition

Setting:

- A network of peers where each node maintains a collection of **documents** (vector of keywords | image features, etc)



Goals:

- **Effectively** query the distributed documents by keywords.
- Consume the less possible network resources.

Search in Unstructured P2P Environments

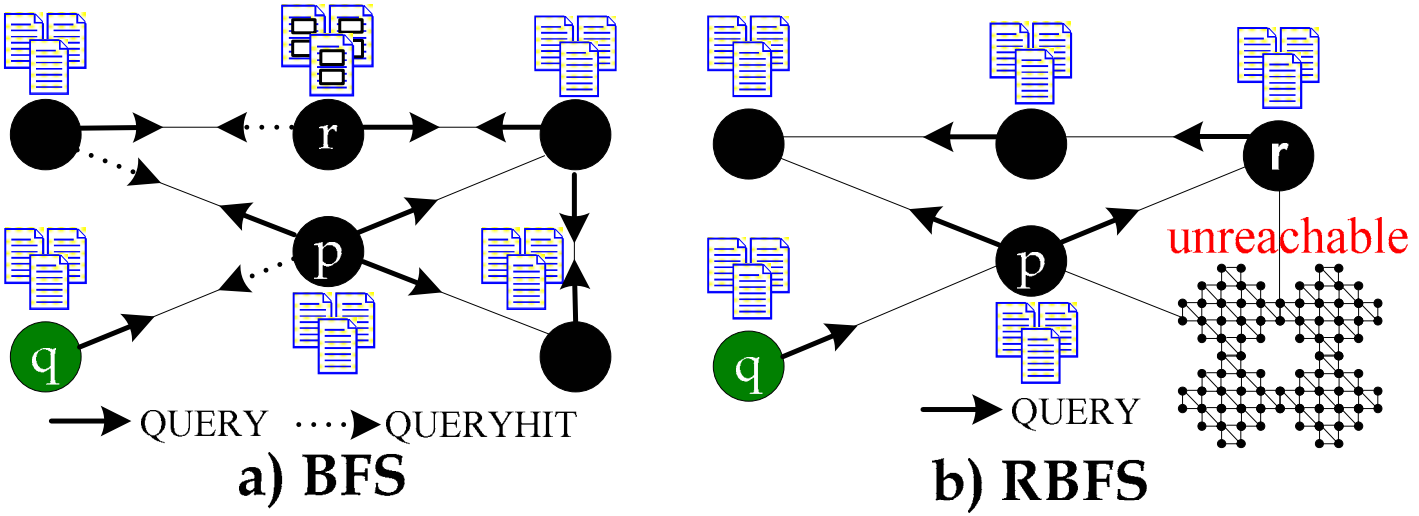
Agnostic Techniques

a) TTL-based Breadth-First-Search (BFS)

- + Each Node forwards the query to all its neighbors.
- - Excessive network and resource consumption.

b) Random BFS (RBFS) [CIKM'02]

- + Each Node forwards the query to a random subset of neighbors.
- - Some important segments may become unreachable.



Search in Unstructured P2P Environments

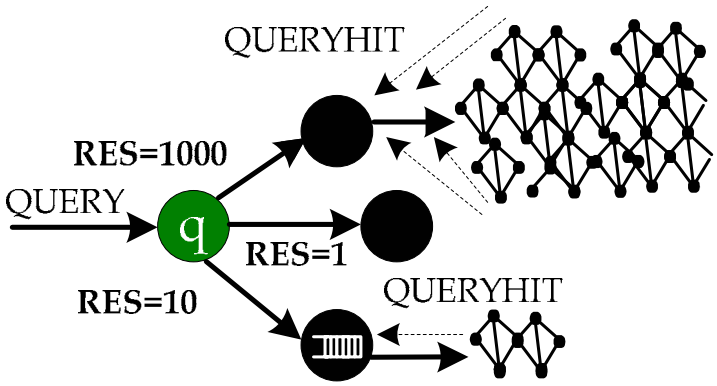
Techniques using Past Statistics

c) Most Results in Past Heuristic (>RES) [IPTPS'02]

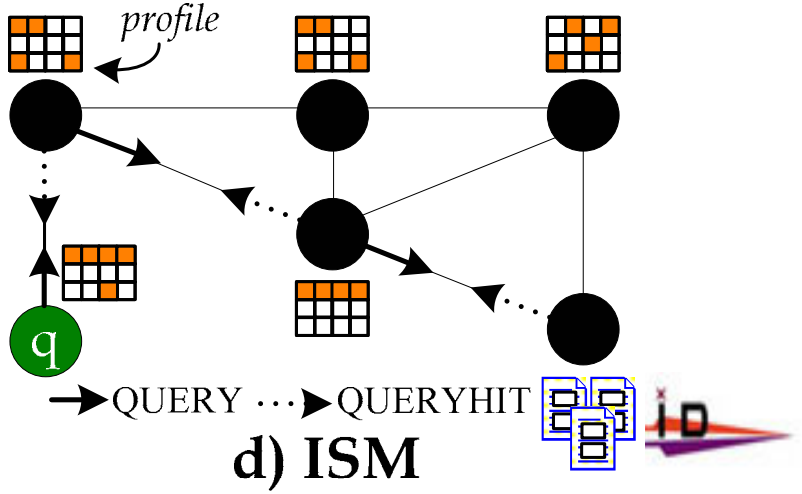
- Query the nodes with the most results in the last K queries
- Usually explores the larger network segments but ….
- … fails to explore the nodes with the most **relevant content**

d) Intelligent Search Mechanism (ISM) [CIKM'02]

- Each Node maintains a query(hit) profile for its neighbors
- Uses the cosine similarity to drive the queries to the results



c) >RES



d) ISM

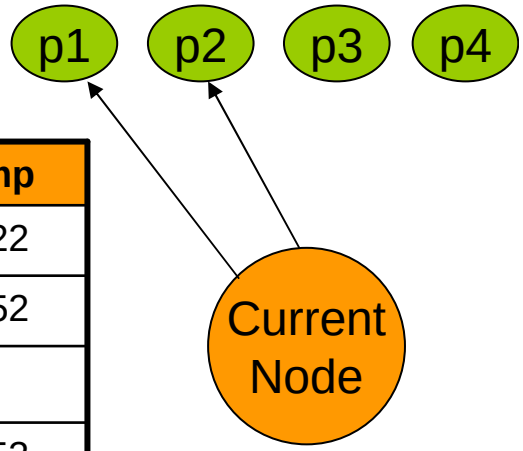
Motivation

- We crawled the Gnutella P2P Network for 5 hours with 17 workstations.
- We analyzed **15,153,524** query messages.
- **Observation:** High locality of specific queries.
- *We try to exploit this property for more efficient searches*

#	Query	Occurrence	%	#	Query	Occurrence	%
1	divx avi	588,146	3,88%	**11**	divx	24,363	0,16%
2	spiderman avi	50,175	0,33%	**12**	spiderman	23,274	0,15%
3	p__ mpg	39,168	0,25%	**13**	xxx avi	22,408	0,14%
4	star wars avi	38,473	0,25%	**14**	capture the light	21,651	0,14%
5	avi	29,911	0,19%	**15**	buffy mpg	20,365	0,13%
6	s__ mpg	27,895	0,18%	**16**	g__ mpg	20,251	0,13%
7	Eminem	27,440	0,18%	**17**	buffy avi	19,874	0,13%
8	eminem mp3	25,693	0,16%	**18**	t__ mpg	19,492	0,12%
9	dvd avi	25,105	0,16%	**19**	seinfeld vivid	18,809	0,12%
10	b___	24,753	0,16%	**20**	xxx mpg	18,686	0,12%

Search in Unstructured P2P Environments

Intelligent Search Mechanism (ISM)



a) Profile mechanism

Query	GUID	Connection & Hits	Timestamp
Olympic Games	G439ID	(peer1,20), (peer4,50),…	100002222
VLDB Athens	F549QL	(peer2,10)	100065652
***	***	***	***
Florida storm	PN329D	*NULL*	100022453

|L|-dim space: {olympic,games,vldb,athens,florida,storm}
e.g. If **q**="athens olympic" => $\vec{\mathbf{q}}$ (vector of q) = [1,0,0,1,0,0]

b) Cosine Similarity – The Similarity Function

$$sim(q, q_i) = cos(q, q_i) = \frac{\sum(\vec{q} * \vec{q_i})}{\sqrt{\sum(\vec{q})^2} * \sqrt{\sum(\vec{q_i})^2}}$$

c) RelevanceRank – Ranking Neighbors by similarity

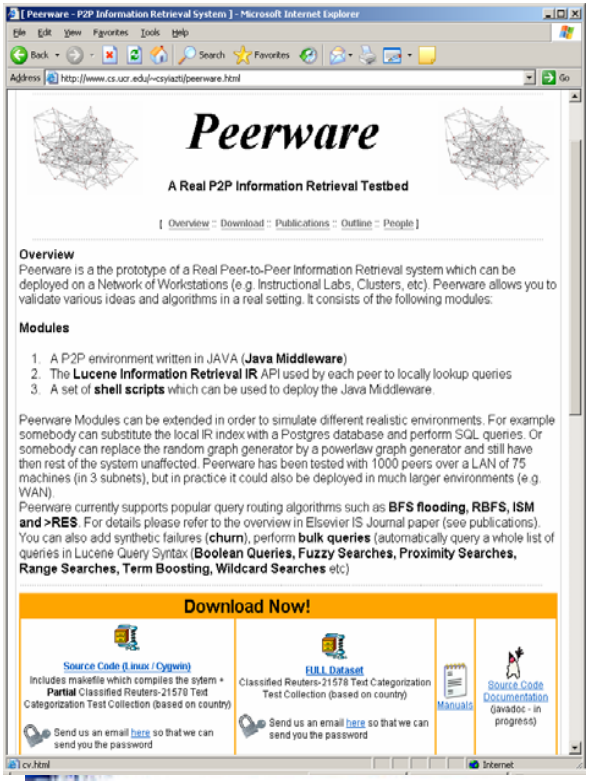
$$RR(peer_i, q) = \sum_{q_j = \text{"Queries answered by } peer_i\text{"}} sim(q_j, q)^{\alpha} * results(q_j)$$

Experimental Evaluation

- We developed a distributed News Agency (useful for Citizen Journalism, Video Sharing) using our open source Peerware system.

  http://www.cs.ucr.edu/~csyiazti/peerware.html

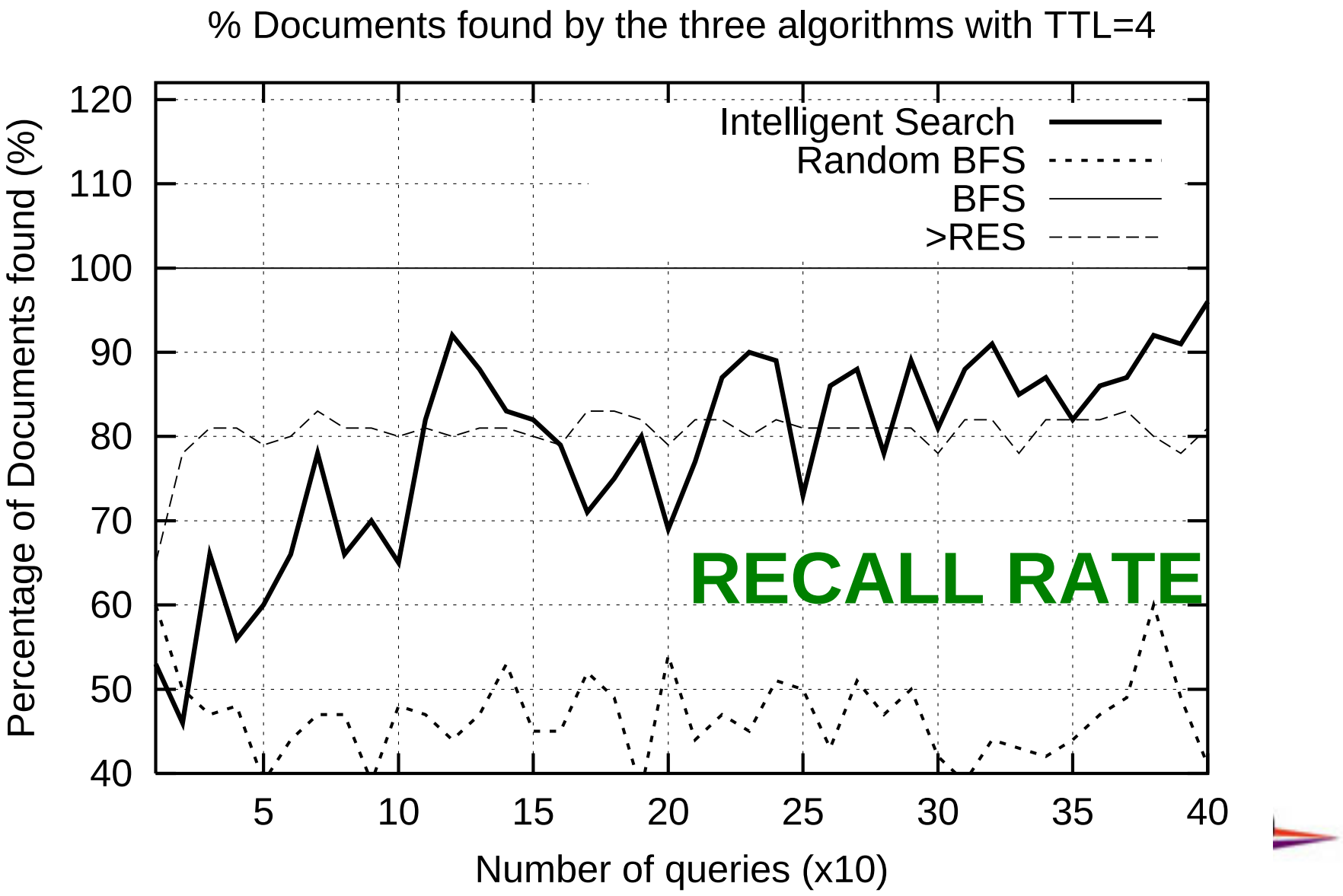
• Each peer utilizes Apache's **Lucene** Information Retrieval system to efficiently organize information locally.



• We run experiments on **75 workstations** (up to 1000 peers connected in a **random topology)** with datasets from **Reuters** and **TREC Los Angeles Times.**

• We compare **Recall Rate** vs. **Num. of messages**

Experimental Evaluation



Experimental Evaluation

Summary of Results

- The ISM achieves in some cases 100% Recall Rate while using 40-50% less Messages and **30-40% less Time** than BFS.
- ISM performs extremely well under failures – high churn rates (10% failure rate still yields 85% recall rate)
- Scales well to large environments (since only local information is utilized)
- Performs best with high locality of queries

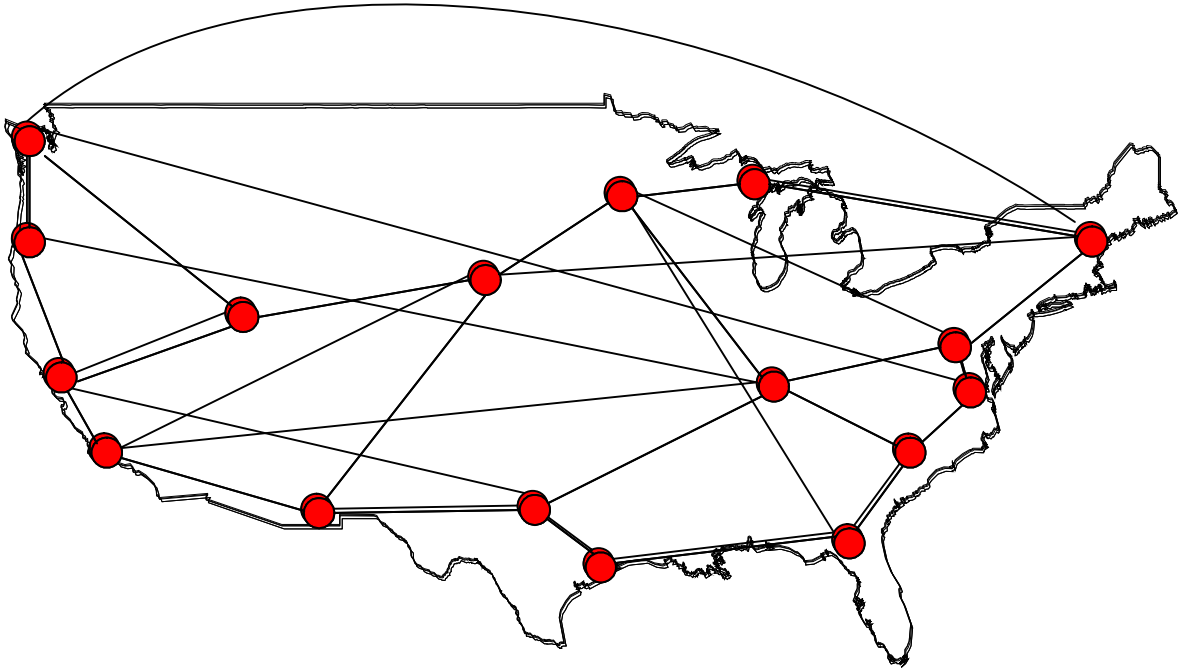
B) Topologically-Aware Overlay Networks

Network Mismatch

- P2P Networks are usually **network-agnostic.**
- **Physical** with **Overlay** Network Mismatch

The Random Overlay Network

The Topologically Aware Network



Network-Efficient Topologies

- The **network mismatch** between the **Physical** and the **Overlay** layer results in high latencies and excessive network resource consumption.
- **Smaller Latency =>** Faster Interaction and Higher Data Transfer rates because of TCP windowing.
- We will discuss the following topologies
  - **i) RANDOM Topology. (Network-Agnostic)**
  - **ii) Short-Long (SL) Topology.**
  - **iii) Binning SL (BinSL) Topology.**
  - **iv) DDNO Topology (used in pFusion)**

(Network-Aware)

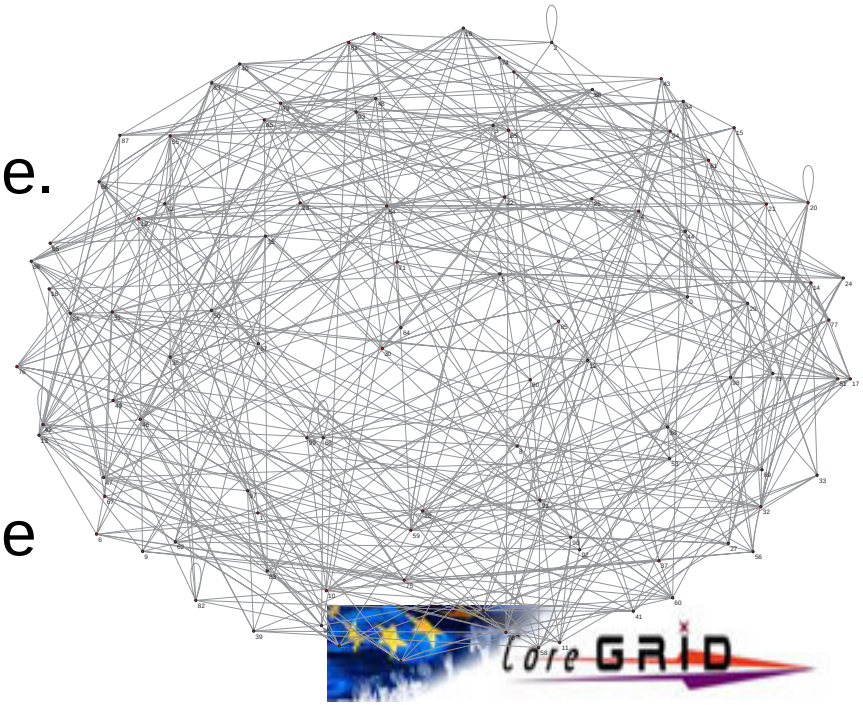
i) Random Topology

a) **Random Topology**

- Each peer randomly connects to k other peers.
- This is the technique used in most systems (such as Gnutella v0.4)

- **Advantages**
  - Simplicity.
  - Needs only Local Knowledge.
  - Leads to connected topologies if *degree* > *logn*
- **Disadvantages**
  - Doesn't take into account the underlying network

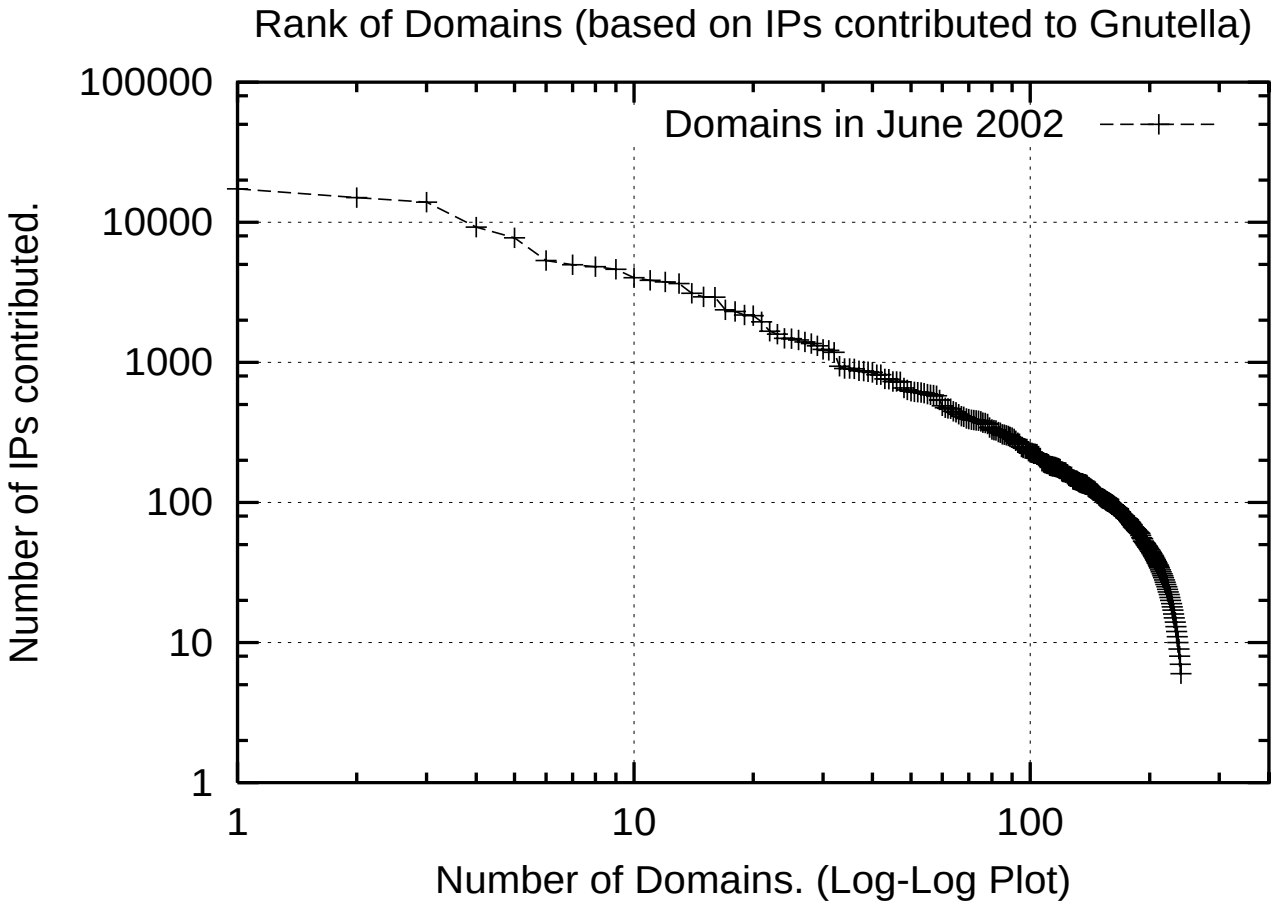
A Random Graph of 100 nodes



DDNO – Distributed Domain Name Order

Motivation

By our analysis of Gnutella (300,000 IPs) we found that 58% of the network belongs to only 20 ISPs



#	Domain	%
1	rr.com	10%
2	aol.com	8%
3	t-dialin.net	6%
4	attbi.com	6%
5	comcast.net	3%

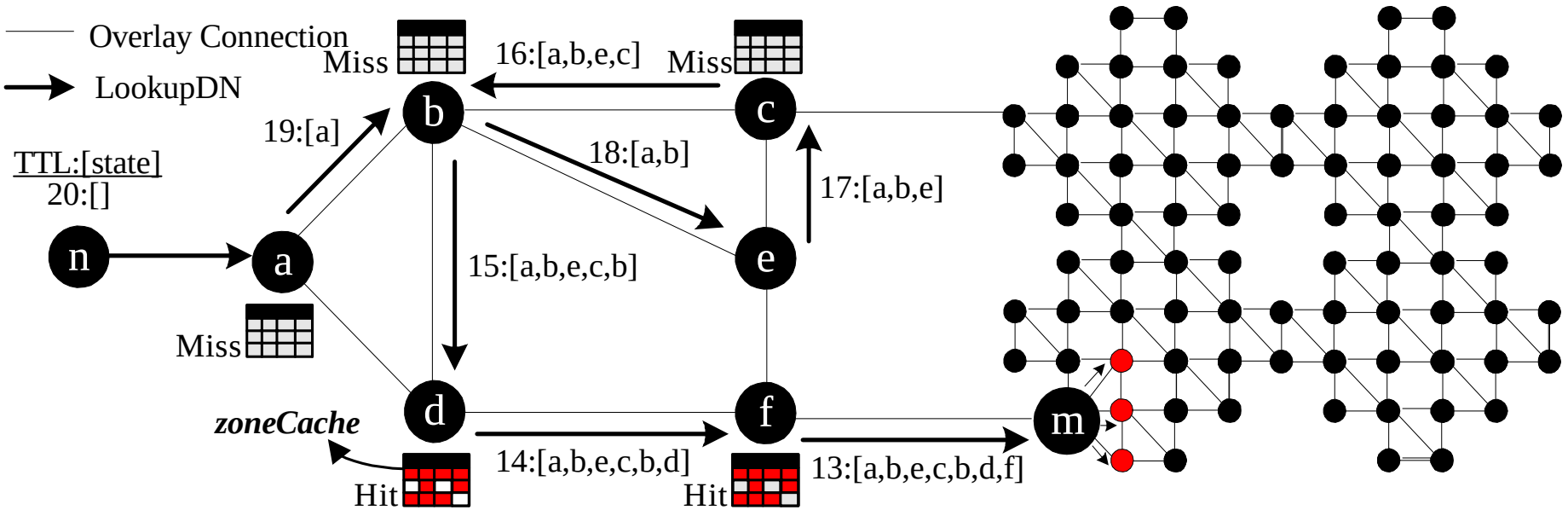
DDNO – Distributed Domain Name Order

Task: Locate nodes in the same domain without any global knowledge

Solution

Naïve Solution: Perform a Random Walk.

Improved Search: Deploy a ZoneCache which tells a node towards which direction to move.



ii) Short Long Topology

b) **Short-Long Topology** [Infocomm'02]

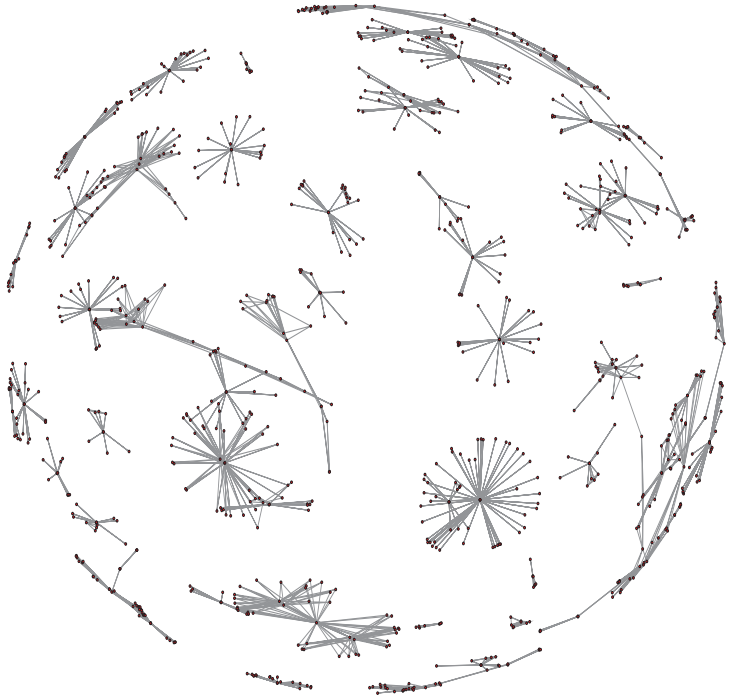
- Build a Global latency adjacency matrix
- Each peer connects to k/2 closest peers **(Short Links)**.
- It then connects to k/2 random peers **(Long Links)**.

- **The construction of the adjacency matrix requires global knowledge.**

  (e.g. each peer pings its neighbors and sends this info to a centralized index)

Note: By choosing only Short Links yields disconnected topologies

A Short Graph of 1000 nodes



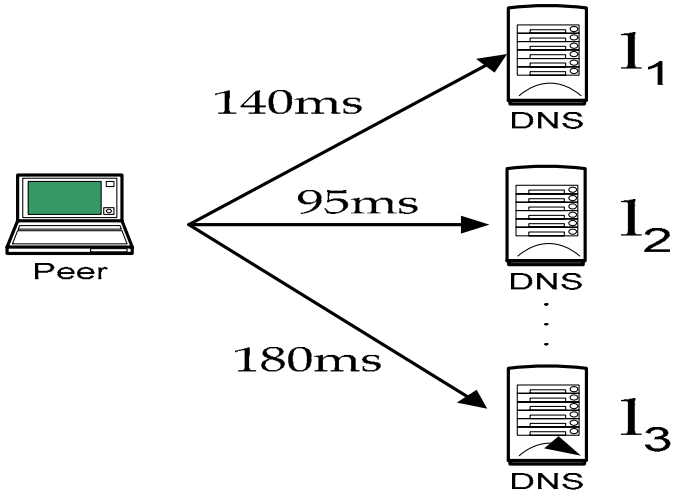
iii) BinSL Topology

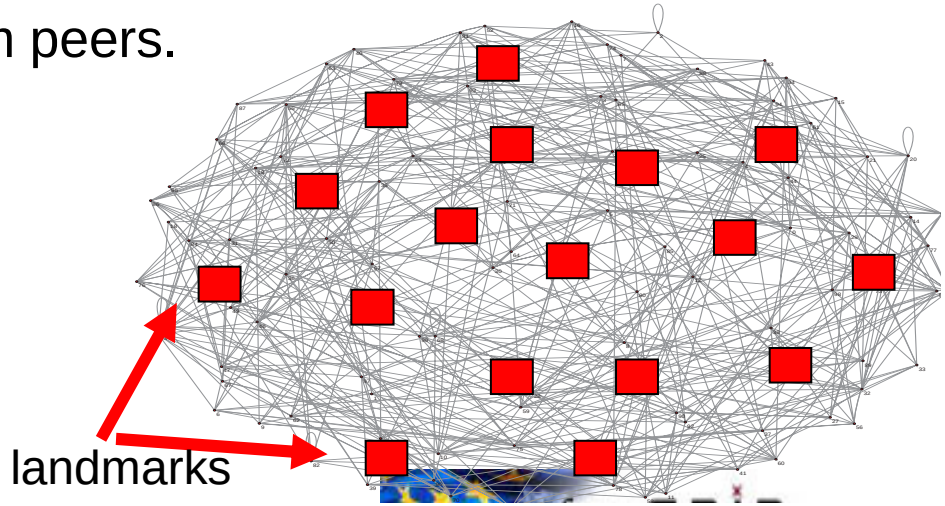
c) Binning SL Topology (BinSL) [Infocomm'02]

Approximate Distances with Distributed Binning

1. Each node calculates the RTT to k well-known landmarks.
   - The numeric ordering of the landmarks defines the ***bin*** of a node.
   - Furthermore latencies are divided into ***level*** ranges. e.g. Level0=[0,100)ms Level1=[100,200)ms , Level3=rest

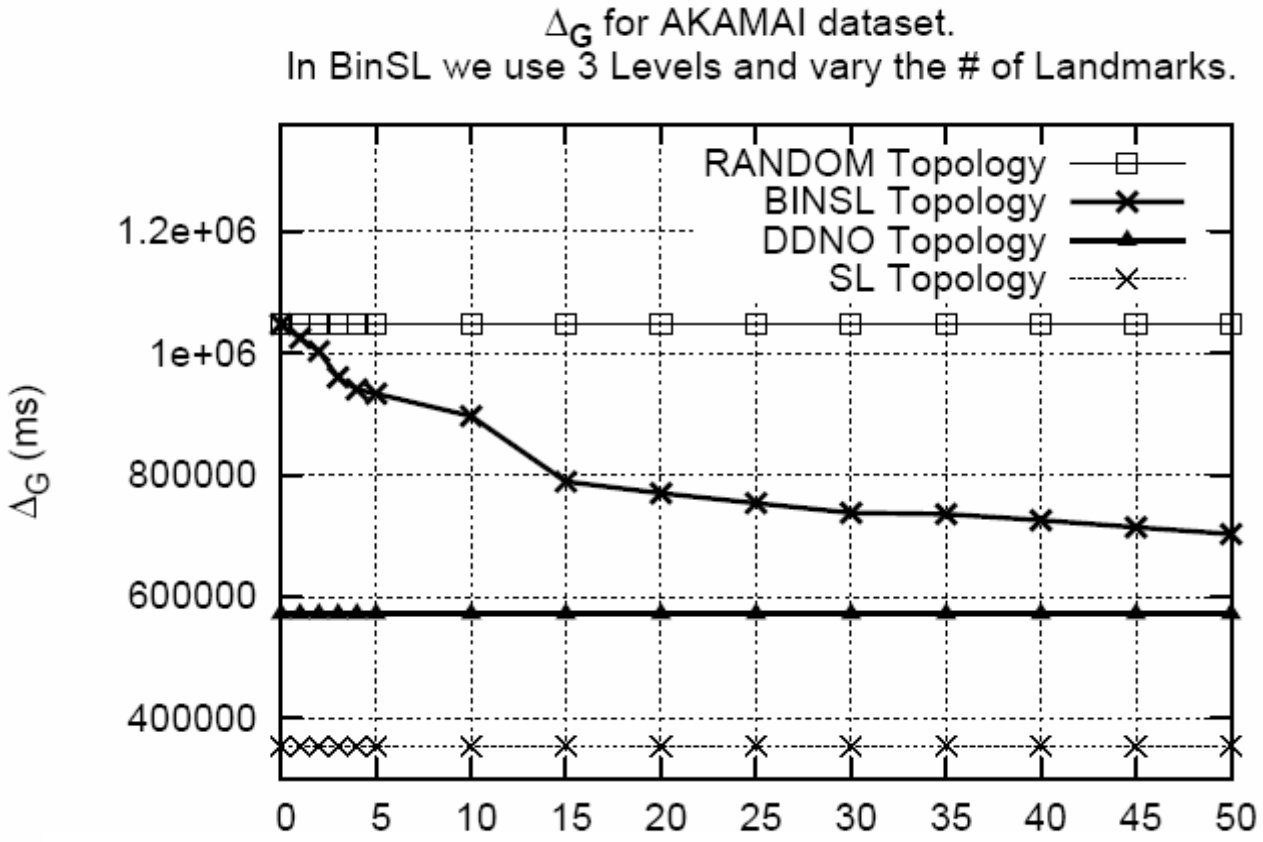
   BinCode = Landmarks:Levels = $l_2 l_1 l_3$:011
2. Each peer then connects to k/2 peers that have the same bin code.
3. It then connects to k/2 random peers.

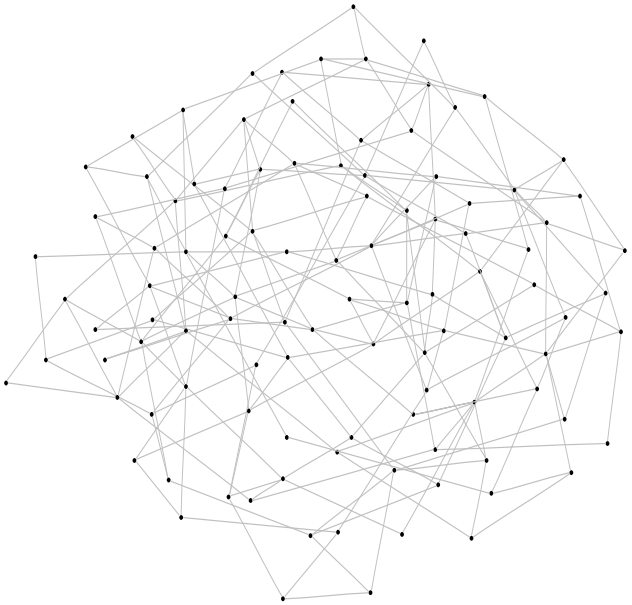




Well-chosen Landmarks

BINSL Drawback: Depends on the Number and Quality of Landmarks



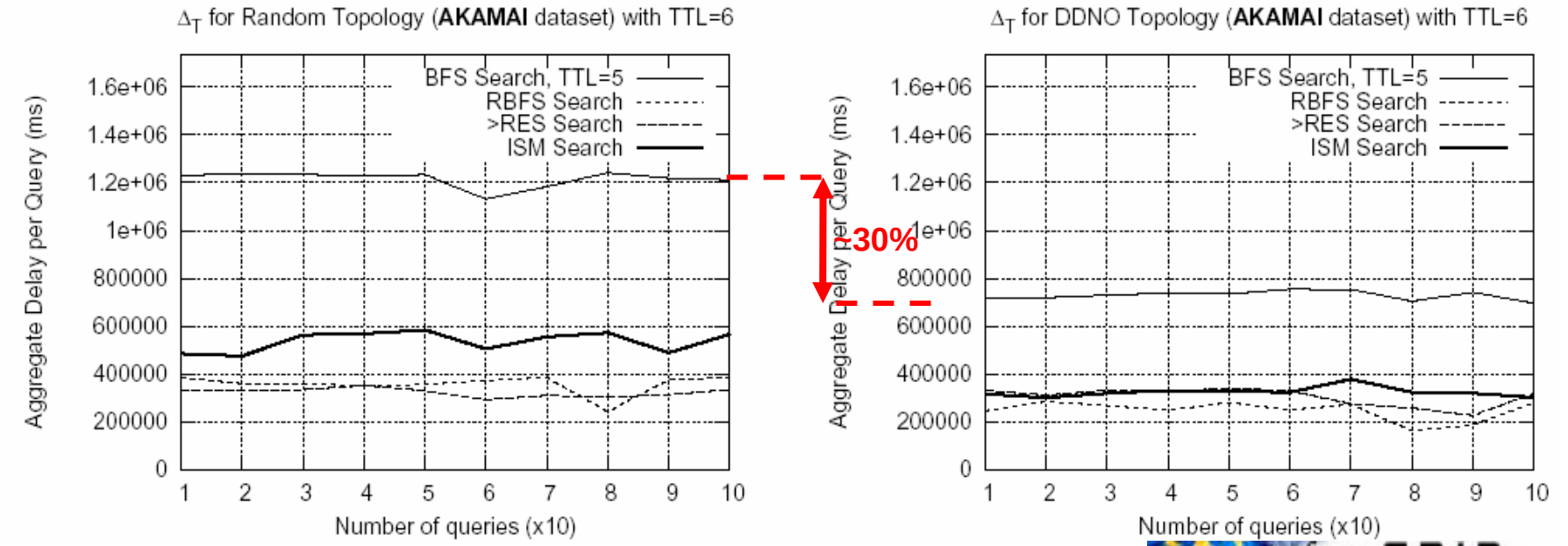


Δ_G : the sum of delays across all the edges of the 1000-node Overlay Graph

The overlay latencies were taken from the AKAMAI CDN

DDNO Advantages

- We perform a Query and measure the delay until the expected answer arrive.
- We observe that a **DDNO** network minimizes this delay for all search methods (BFS, RBFS, >RES and ISM) by 30% over **RANDOM**



RANDOM GRAPH

DDNO GRAPH

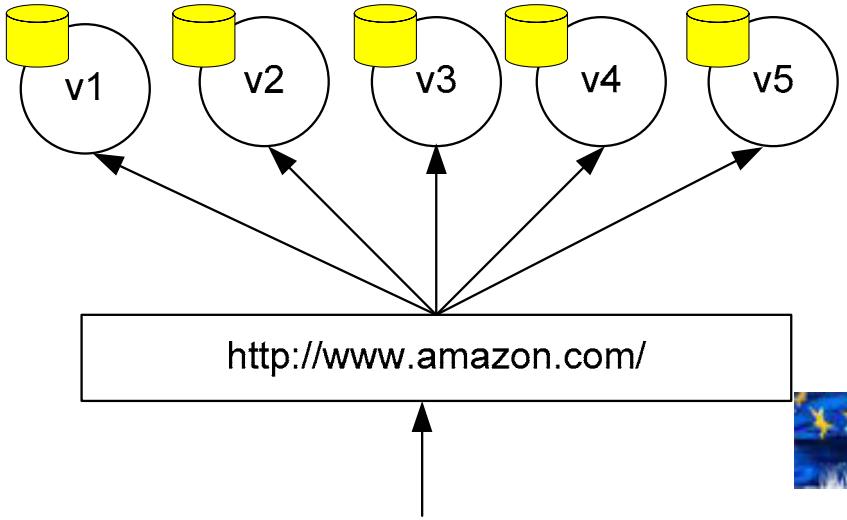
Other Research Activities

Distributed Top-K Query Processing

(in collaboration with IBM Almaden & AT&T Research and Univ. of Toronto)

Problem Example

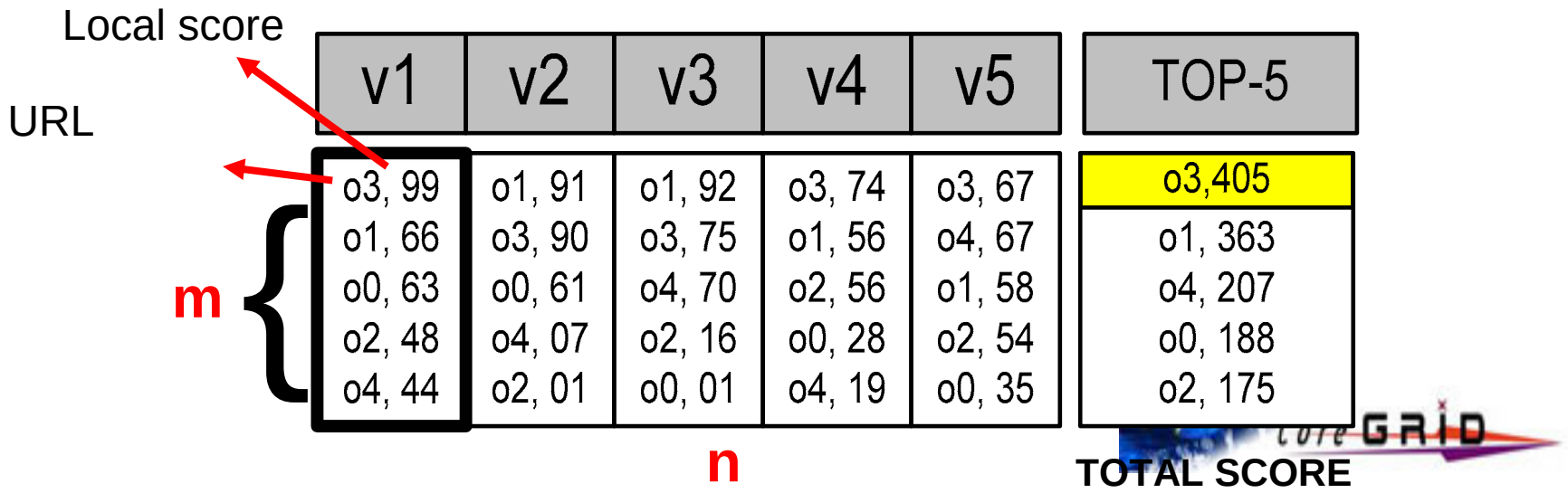
- Assume that we have a cluster of **n=5 webservers.**
- Each server maintains locally the same **m=5 webpages.**
- When a web page is accessed by a client, a server increases a local **hit counter** by one.



Distributed Top-K Query Processing

Problem Example (cont'd)

- ***TOP-1 Query:*** *"Which Webpage has the highest number of hits across all servers (i.e. highest* ***Score(o_i))****?"*
- ***Score(o_i)*** *can only be calculated if we combine the hit count from all 5 servers.*



Other Research I

Distributed Top-K Query Processing

Publications:

- **"The Threshold Join Algorithm for Top-k Queries in Distributed Sensor Networks",**

  D. Zeinalipour-Yazti, Z. Vagena, D. Gunopulos, V. Kalogeraki, V. Tsotras, M. Vlachos, N. Koudas, D. Srivastava In **ACM DMSN (VLDB'2005**), Trondheim, Norway, pp. 61-66, 2005.

- **"Data Acquision in Sensor Networks with Large Memories“,**

  D. Zeinalipour-Yazti, S. Neema, D. Gunopulos, V. Kalogeraki and W. Najjar,, IEEE Intl. Workshop on Networking Meets Databases **IEEE NetDB (ICDE'2005),** Tokyo, Japan, 2005.

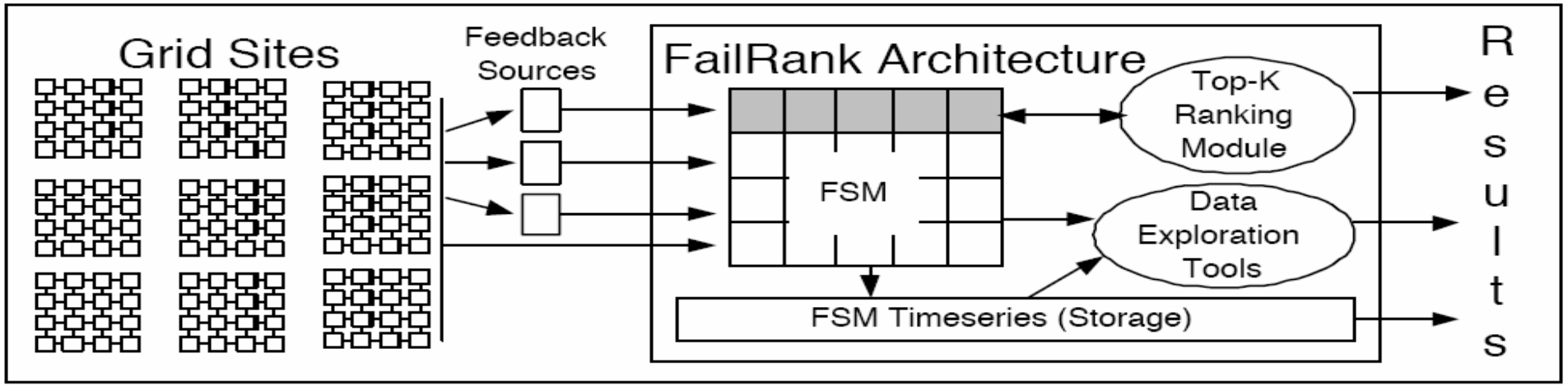
- **"Distributed Spatio-Temporal Similarity Search"**

  D. Zeinalipour-Yazti, S. Lin, D. Gunopulos, ACM 15th Conference on Information and Knowledge Management, (**ACM CIKM 2006**), November 6-11, Arlington, VA, USA,

Other Research II

FailRank: Failure Management in Grids

- **Motivation:** Studies have shown that a very large percentage of scheduled jobs fail on EGEE.
- **Task:** Design a Failure Management Module that will allow Grid Resource Brokers to divert jobs away from failing Grid Sites
- **Core Idea:** Continuously fuse together meta-information that is available for Job Queues trough monitoring services web sites, LDAP, etc. then rank this information.



Other Research II

FailRank: Failure Management in Grids

- **Status:**
  - We collected a 4GB trace for 2,500 queues (1 month) from a variety of sources (**S**ite**F**unctional**T**ests, LDAP queries on Information Service, etc)
  - We are currently analyzing this trace to quantify failures (i.e., associate grid status state with failures).

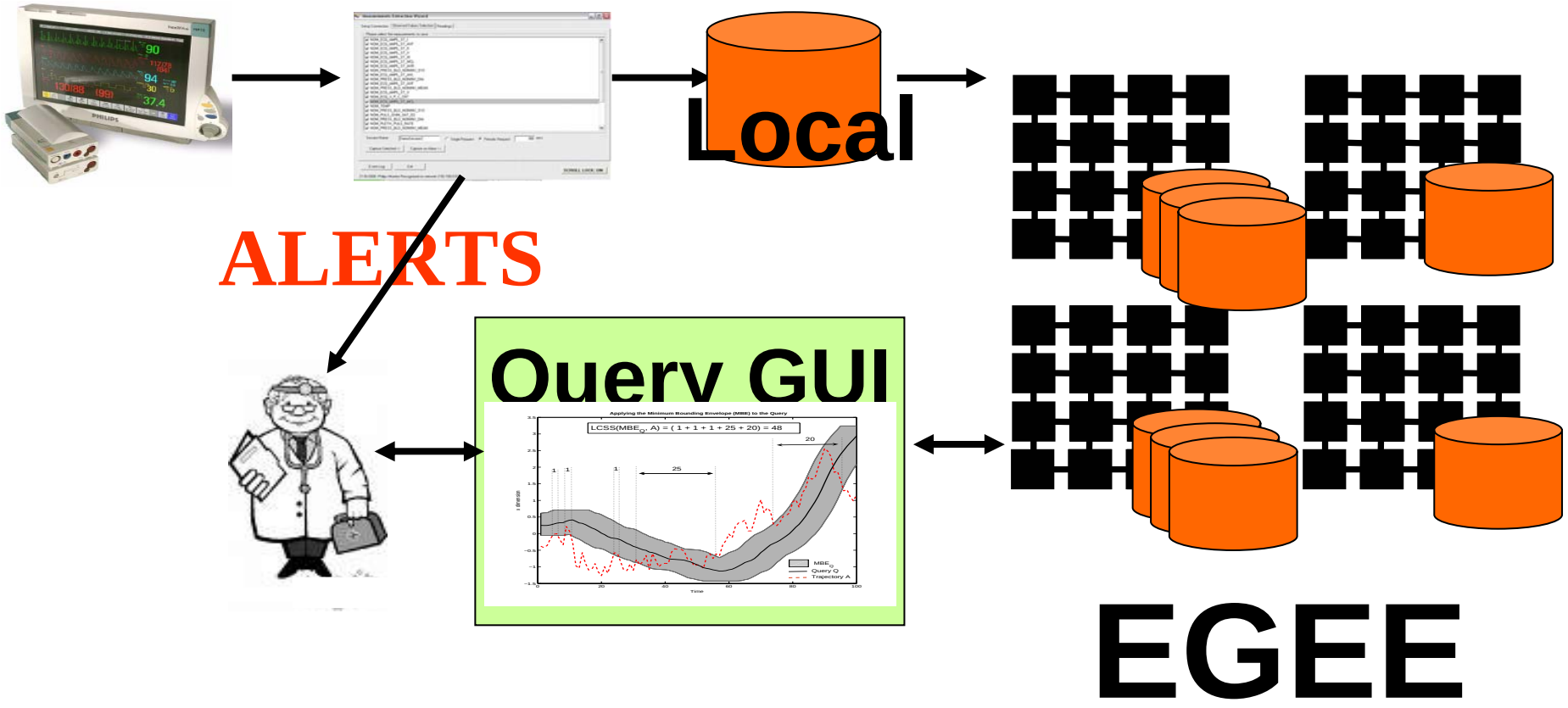
- **Publications:**
  - **"Managing failures in a Grid System using FailRank."** D. Zeinalipour-Yazti, K. Neokleous, C. Georgiou, M.D. Dikaiakos, Technical Report TR-2006-04, Department of Computer Science, University of Cyprus, September 2006

  - **"Monitoring and Ranking of Grid Failures using FailRank."** D. Zeinalipour-Yazti, K. Neocleous, C. Georgiou, M.D. Dikaiakos, EGEE '06 Conference (**poster session**), Geneva, September 26, 2006.

Other Research III

ICGrid: Intensive Care Grid

ICGrid is a distributed platform that enables the integration, correlation and retrieval of clinically interesting episodes across Intensive Care Units



Other Research III

ICGrid: Intensive Care Grid

- **Demonstrations and Posters:**
  - **Poster:** "ICGrid: Intensive Care Grid." D. Zeinalipour-Yazti, M.D. Dikaiakos, M. Papa, T. Kyprianou, G. Panayi, EGEE '06 Conference (poster session), Geneva, September 26, 2006.
  - **Demo: "ICGrid: Intensive Care Grid."** D. Zeinalipour-Yazti, H. Gjermundrod, M. D. Dikaiakos, G. Panayi and Th. Kyprianou, CoreGRID Industrial Confernece (part of the GRIDS@WORK series of events), Sophia-Antipolis, Nov. 30-Dec. 1, 2006 (best demonstration award).
  - **Invited Demo: "ICGrid: Intensive Care Grid."** D. Zeinalipour-Yazti, H. Gjermundrod, M. D. Dikaiakos, G. Panayi and Th. Kyprianou, IST Conference, Helsinki, Finland, Nov. 17, 2006.

Other Research IV

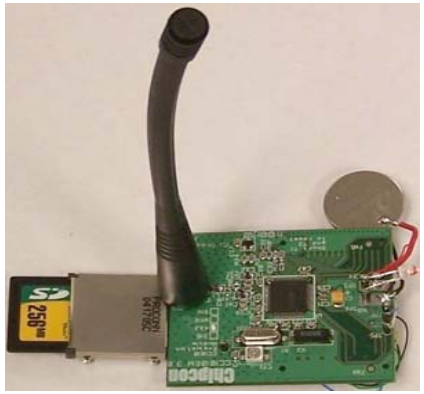
Data Management in Sensor Networks

(in collaboration with University of California, Riverside)

- **Motivation:** Transmitting over the radio is the most expensive function in a Wireless Sensor Network.

- **Core Idea** (The **In-Situ Storage** Paradigm):
  - Store the data locally at each sensor
  - Access this information with On-demand Queries rather than continuously.

- **Main Contribution**
  - **The Ri**verside**Se**nsor **System**
  - **Provides index structures for giga-scale storage and retrieval in sensor systems**



Other Research IV

Data Management in Sensor Networks

We designed the Microhash Index Structure that enables efficient access to values stored on Flash Media (the most prevalent storage media of sensor systems).

Publications:

- ***MicroHash: An Efficient Index Structure for Flash-Based Sensor Devices",***

  D. Zeinalipour-Yazti, S. Lin, V. Kalogeraki, D. Gunopulos and W. Najjar, The 4th **USENIX** Conference on File and Storage Technologies (**FAST'05**), 2005.

- **"Efficient Indexing Data Structures for Flash-Based Sensor Devices",**

  S. Lin, D. Zeinalipour-Yazti, V. Kalogeraki, D. Gunopulos, W. Najjar **ACM Transactions on Storage (TOS),** ACM Press, In Press, 2006.

Content-Based Search in Internet-Scale Peer-to-Peer Systems

by

Demetris Zeinalipour

Visiting Lecturer

Department of Computer Science

University of Cyprus

Thank you!
Happy New Year!

Thursday, December 28th, 2006
Royal Institute of Technology (KTH)
Stockholm, Sweden

http://www.cs.ucy.ac.cy/~dzeina/

CoreGRID