
Information Retrieval and Filtering over Self-Organising Digital Libraries

Paraskevi Raftopoulou^{1,2}, Euripides G.M. Petrakis²,
Christos Tryfonopoulos¹, and Gerhard Weikum¹



mpi

¹Max-Planck Institute for Informatics, Saarbruecken, Germany
<http://www.mpi-inf.mpg.de/>



²Technical University of Crete, Chania, Greece
<http://www.intelligence.tuc.gr/>

Outline

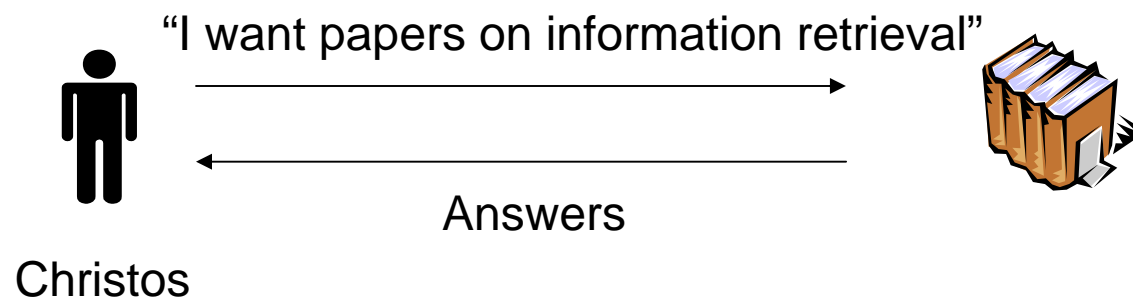
- Motivating scenario
- Background
- iClusterDL
 - Architecture
 - Protocols
- Experimental evaluation
- Related work & outlook



Motivating scenario

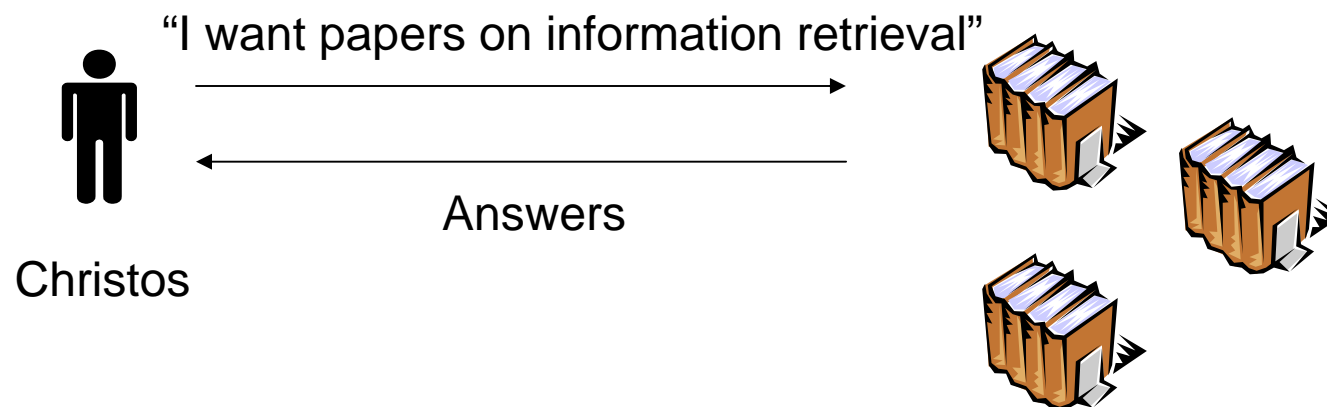
Motivating scenario

- Christos needs papers on information retrieval



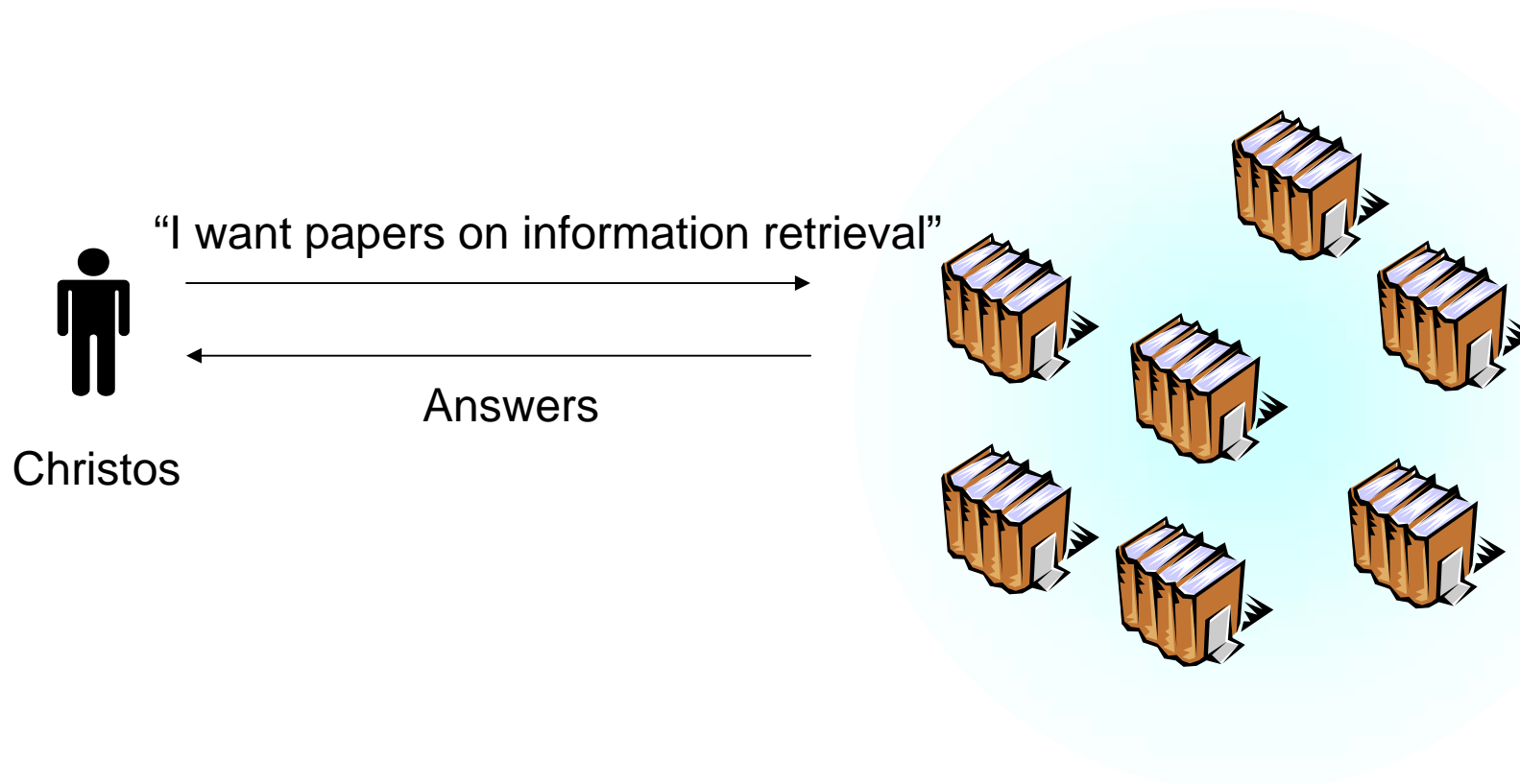
Motivating scenario

- Christos needs papers on information retrieval



Motivating scenario

- Christos needs papers on information retrieval



Motivating scenario

- There are lots of DLs out there!
 - Why ask one or a few, when you could ask thousands?
 - Goal: Distributed resource sharing
- Framework to provide IR and IF functionality on top of SONs
- Integrate DLs, publishers and other networks seamlessly and with minimum effort
- Speed-up query processing



Background information

Background: IR vs IF

■ *IR scenario:*

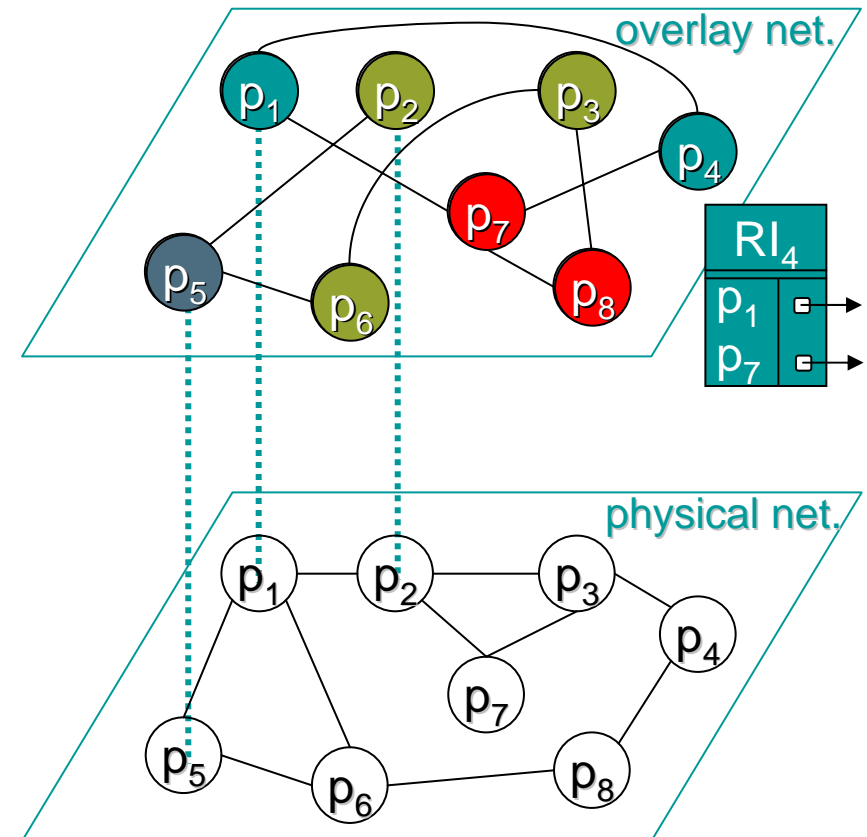
- ❑ A user poses an **one-time query** “I want papers on information retrieval”.
- ❑ The system returns a list of pointers to matching resources (or the actual resources).

■ *IF (or pub/sub or information dissemination) scenario:*

- ❑ A user posts a **continuous query** to receive a notification when a paper on “information retrieval” is published.
- ❑ The system notifies the subscriber with a pointer to the matching resources (or the actual resources).

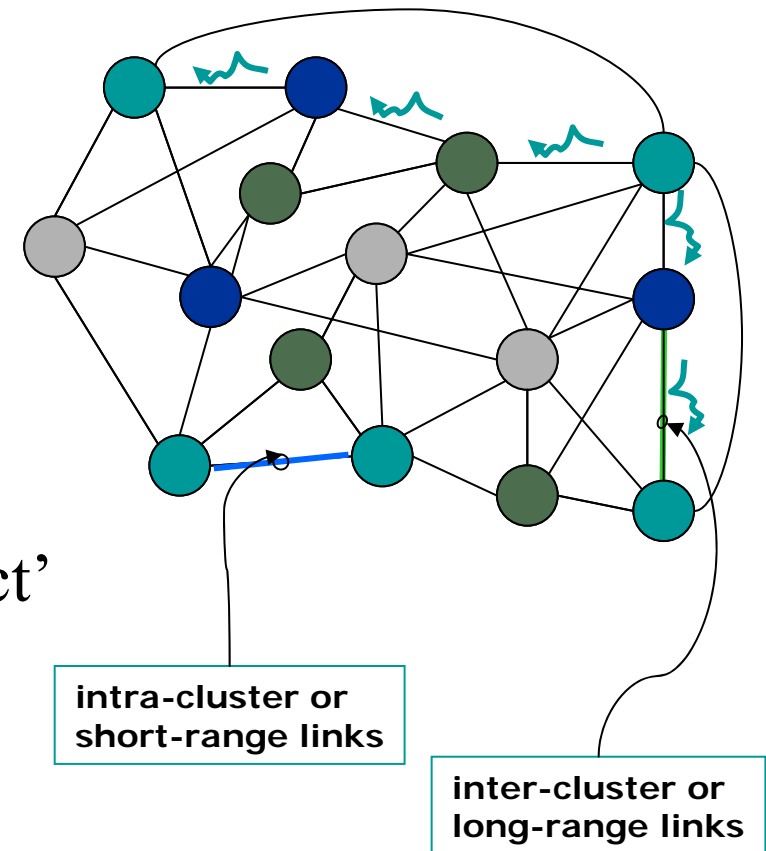
Background: SONs

- **Virtually** connected peers
- Routing indices with links to other peers
- Peers connected to each other are called **neighbors**
- Provide **semantic** (and social) information about peers
- **Self-organising** overlay networks
- Support **rich** data models and **expressive** query languages



Background: Rewiring strategies

- Techniques for **self-organising** peers:
 - **abandon old** connections and **create new** ones
 - **periodic** process
- Inspired by the ‘small world effect’
 - reach anybody in a **small number** of routing hops





iClusterDL architecture

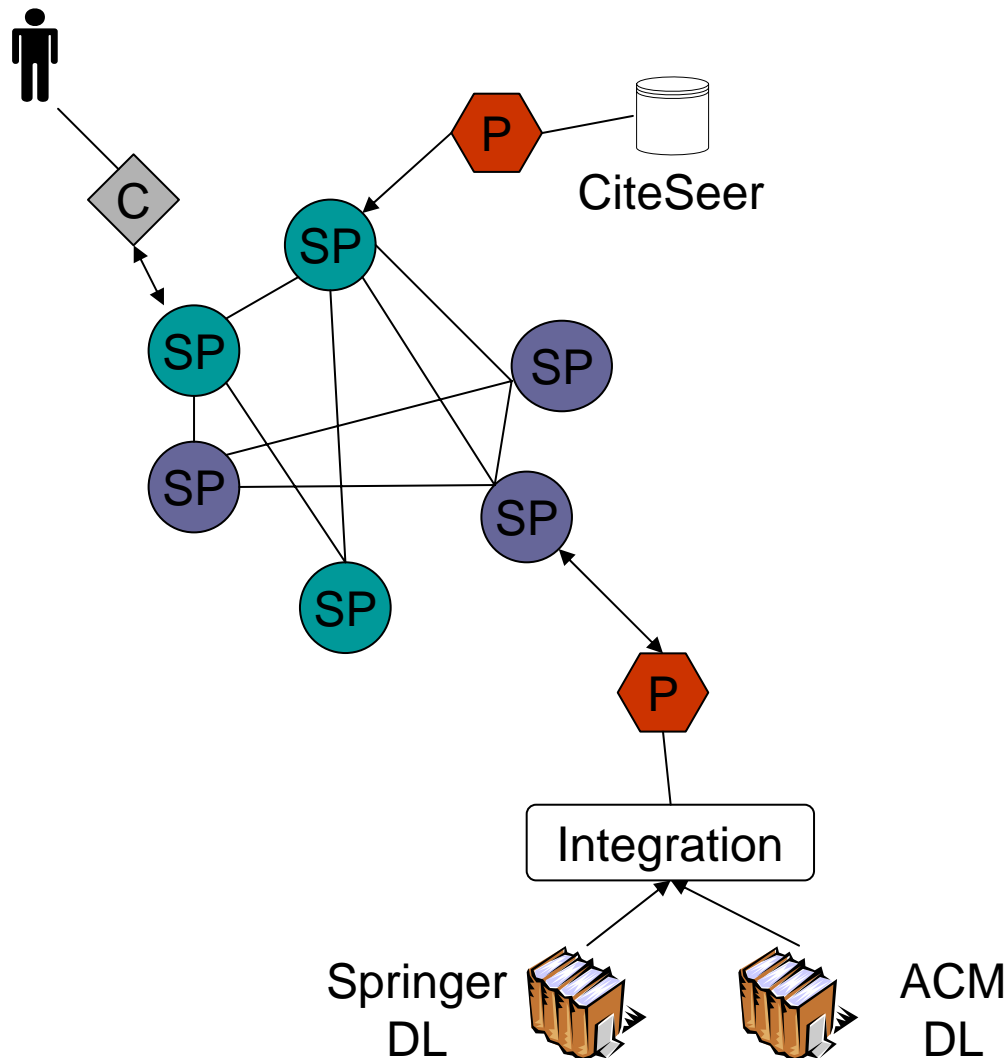
iClusterDL basics

- (i) intelligent + (Cluster) clustering + (DL) digital libraries = iClusterDL

Contributions:

- Architecture and protocols to support **both IR and IF**
 - 2-level hierarchical (super-peer) P2P network
 - seamless and easy integration of DLs, scalable
- Self-organising DLs based on **SONs**
 - support rich query models
 - benefits from loosely-connected peers

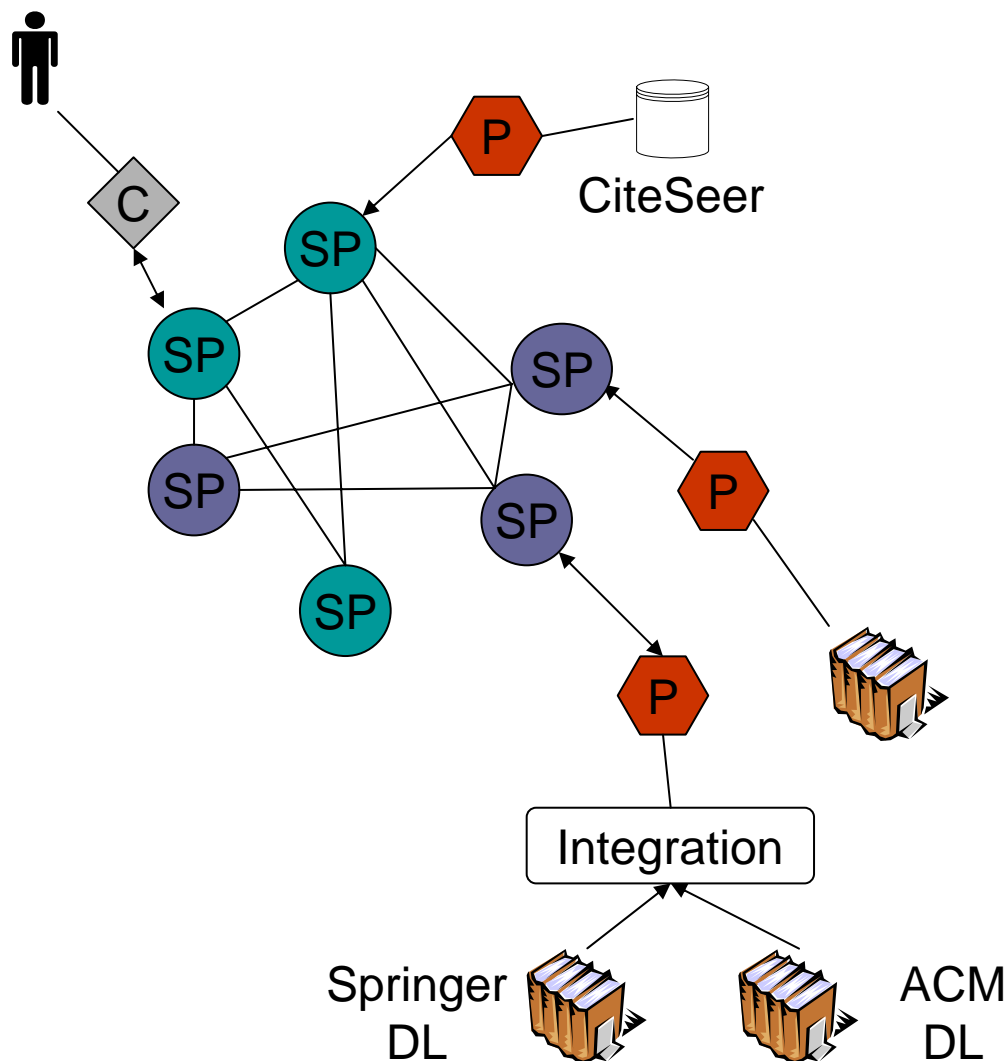
iClusterDL Architecture



SP Super-peer

- Forms message routing layer
- Runs a rewiring protocol
- Serves clients and providers
 - stores cont. queries
 - stores resource publications
 - answers one-time queries
 - creates notifications
 - stores notifications

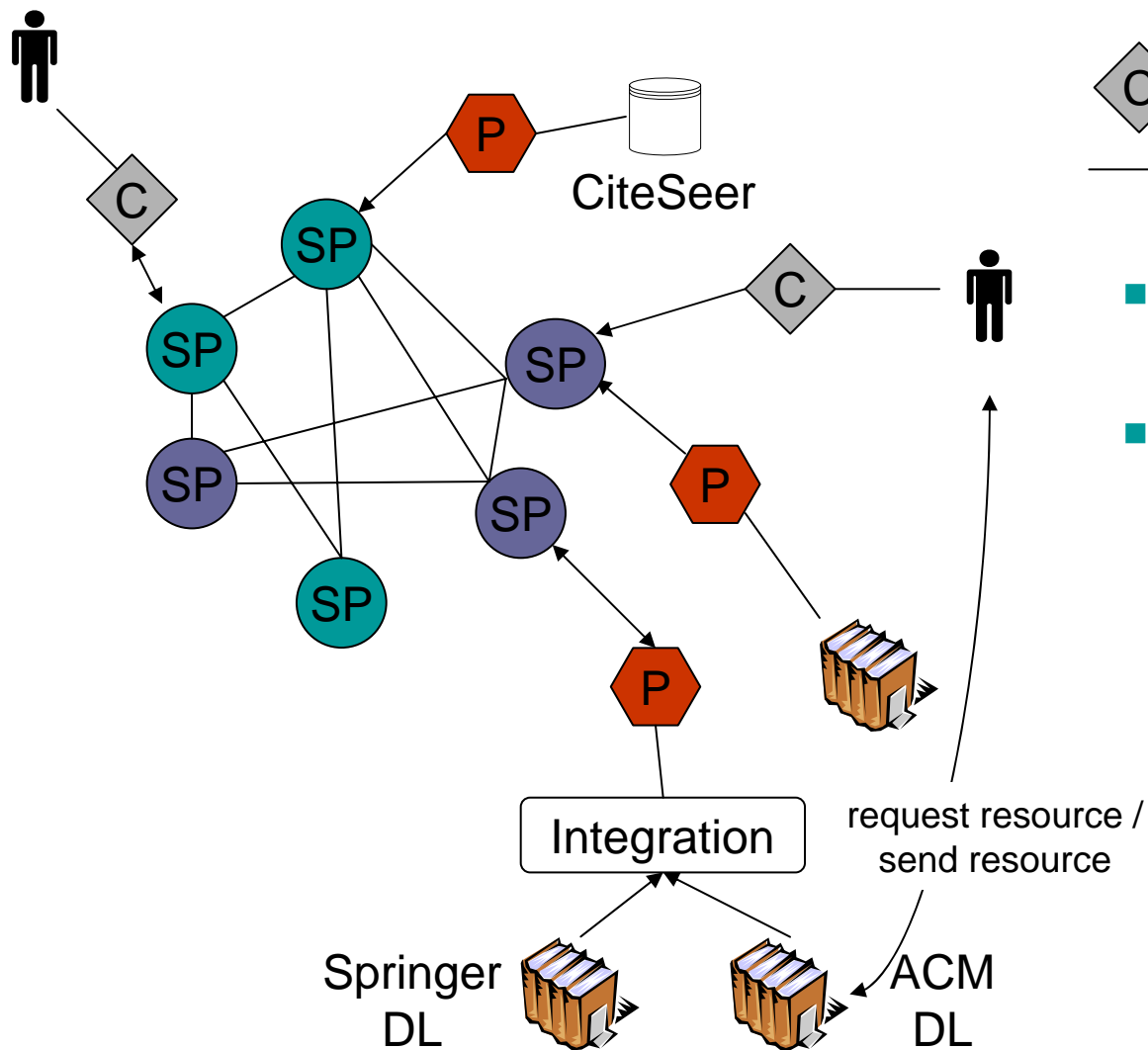
iClusterDL Architecture



P Provider

- Implemented by information sources
- Used to expose source's contents
- Connects to iClusterDL network through a super-peer

iClusterDL Architecture



- Connects to iClusterDL network through a super-peer
- Information consumers:
 - pose one-time queries
 - receive answers
 - subscribe to resource publications
 - receive notifications

iClusterDL Protocols

- Super-peer join/leave
- Super-peer rewiring
- Client join (first time only)
- Client connect/disconnect
- Resource publication/indexing/removal/update
- One-time query processing
- Continuous query processing
- Notification delivery (client online or offline)

Super-peer protocols

- Basic idea: **Organise** super-peers in SONs. Make sure that **similar** super-peers are clustered together.
- Two levels of clustering:
 - A **provider peer** clusters its documents and uses its **interests** to join the network.
 - A **super-peer** uses the interests of its providers to identify itself in the network and find other **similar** super-peers.

Super-peer rewiring

A super-peer s

1. computes its **intra-cluster similarity**
(average similarity with its short-range links)
 2. initiates rewiring if similarity $<$ threshold θ
 3. sends a **message** (msg) with its **interest** to m neighbors
- All super-peers receiving msg **append their interest** and **forward** msg to m neighbors
 - The message is **sent back** to s when $TTL = 0$

IR protocols

- Basic idea: **Index** information in the SON. Make sure one-time queries **meet** similar publications.

- Two levels of indexing:
 - **Global (among all super-peers)**: Use a self-organising protocol.
 - **Local (at each super-peer)**: Use a local index appropriate for the publication language.

One-time query processing

A super-peer s

1. **compares** q against its interests & selects the interest int **most similar** to q
 2. if similarity \geq threshold θ
 - forwards a message (msg) including q to all its **short-range** links
 - sends q to all **similar providers** stored in its provider table
 3. if similarity $<$ threshold θ forwards msg to the **m of its neighbors** most similar to q
- All super-peers receiving msg do the **same process**
 - The message is forwarded until $TTL = 0$



Experimental evaluation

Experimental Evaluation

- Evaluated the protocols under different parameters:

- Data corpus
- Similarity threshold
- Query TTL

- Looked into the:

- Network traffic
- Recall

- ✓ OHSUMED TREC
30,000 medical articles
10 categories
- ✓ TREC-6
556,000 documents
100 categories

Experimental Evaluation

■ Evaluated the protocols under different

parameters:

- ❑ Data corpus
- ❑ Similarity threshold
- ❑ Query TTL

- ✓ the start of the rewiring is randomly chosen from the time interval $[0, 4K]$
- ✓ the periodicity is randomly selected from a normal distribution of $2K$

■ Looked into the:

- ❑ Network traffic
- ❑ Recall

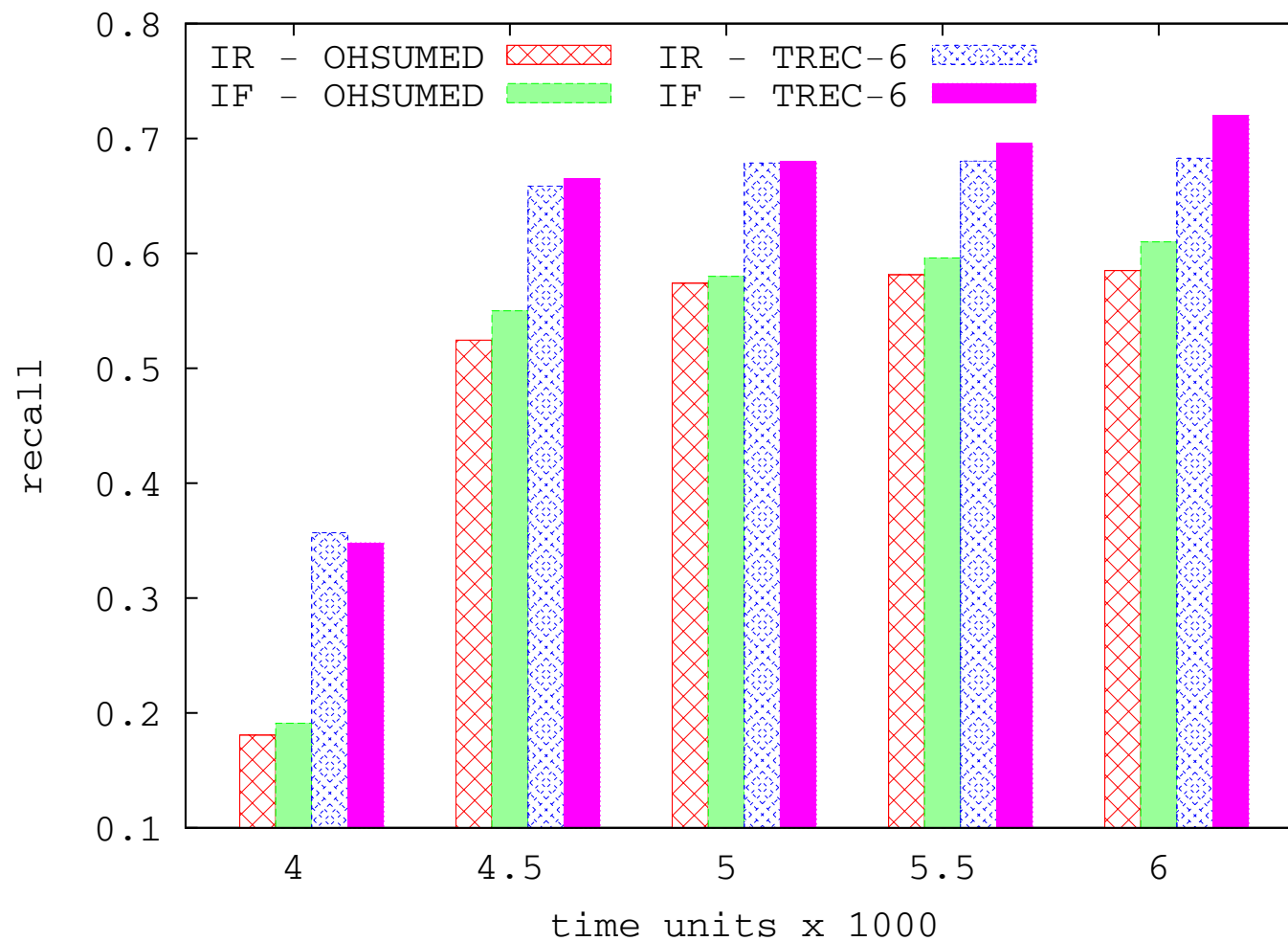
Experimental Evaluation

- Evaluated the protocols under different parameters:
 - Data corpus
 - Similarity threshold
 - Query TTL
- Looked into the:
 - Network traffic
 - Recall

Parameter	Symbol	Value
super-peers	N	2,000
short-range links	s	8
long-range links	l	4
similarity threshold	θ	0.9
rewiring TTL	τ_R	4
fixed forwarding TTL	τ_f	6
broadcast TTL	τ_b	2
message fanout	m	2

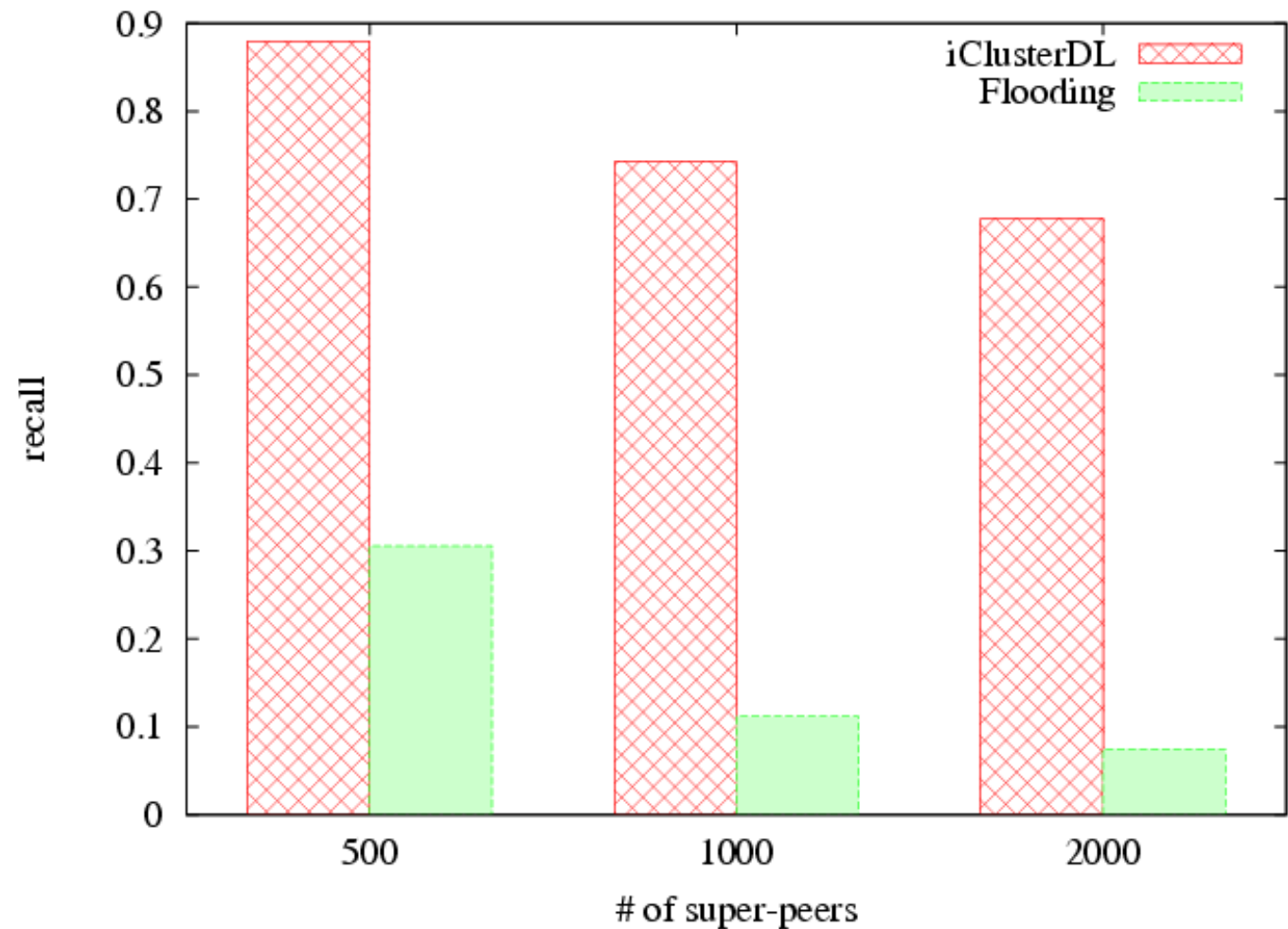
Experimental Evaluation

Recall for IR and IF



Experimental Evaluation

Recall for iClusterDL and Flooding
using the same number of messages





Related work and outlook

Related Work

- Semantic Overlay Networks
 - Initial approaches include:
[KJ04], [SMZ03], [PMW07]
 - Based on the idea of small-world networks:
[Smi04], [LLS04], [VSI06], DESENT

- IR and IF in Digital Libraries
 - Content-based retrieval:
[LC03], OverCite
 - Support both IR & IF functionality:
P2PDIET, LibraRing, MinervaDL

Contributions summarised

- First architecture to
 - **unify** IR & IF functionality in **SONs**
 - apply **SONs** to **Digital Library** domain to support scalability

- An architecture that is
 - **automatic**: requires no intervention
 - **general**: works for any type of data
 - **adaptive**: adjusts to changes of DL contents
 - **efficient**: offers fast query processing
 - **accurate**: achieves high recall

Open Problems

- The effect of different system parameters on:
 - clustering performance
 - retrieval performance
- Dynamic peer content
- Peer churn

Acknowledgements - Funding

- EU project Aeolus
- Heraclitus
(Greek Government PhD Fellowship Program)