On Investigating Scalability and Robustness in a Self-organizing Retrieval System

Jan Sedmidubsky
Vlastislav Dohnal
Pavel Zezula

Faculty of Informatics
Masaryk University
Brno, Czech Republic

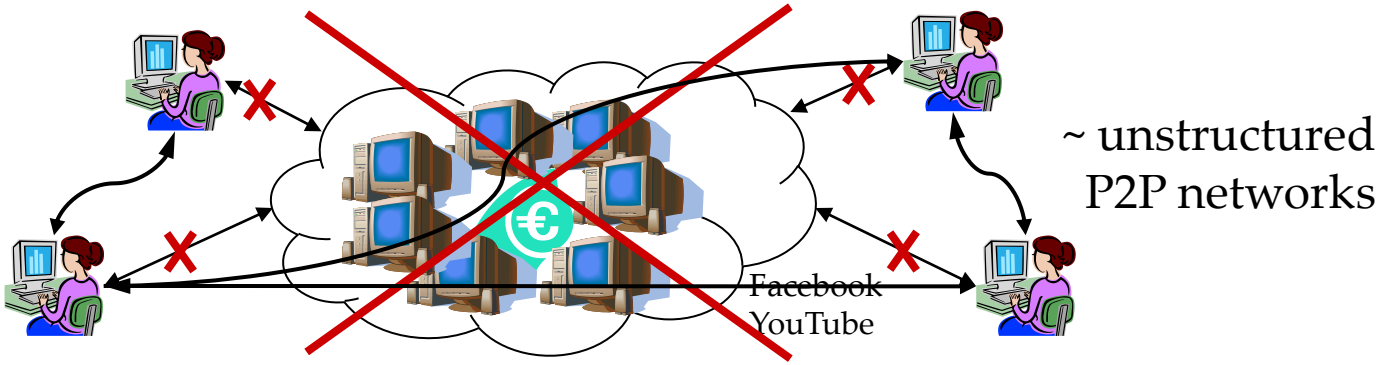
Outline

- Motivation
- Metric Social Network
  - Architecture
  - Query Routing
- Experimental Trials
  - Scalability
  - Adaptability
  - Robustness
- Conclusions

Motivation

- Digital data explosion
  - 100 million new photos uploaded to Facebook everyday
  - 30 hours of videos uploaded to YouTube every minute

⇒ data must be efficiently stored, shared, and searched



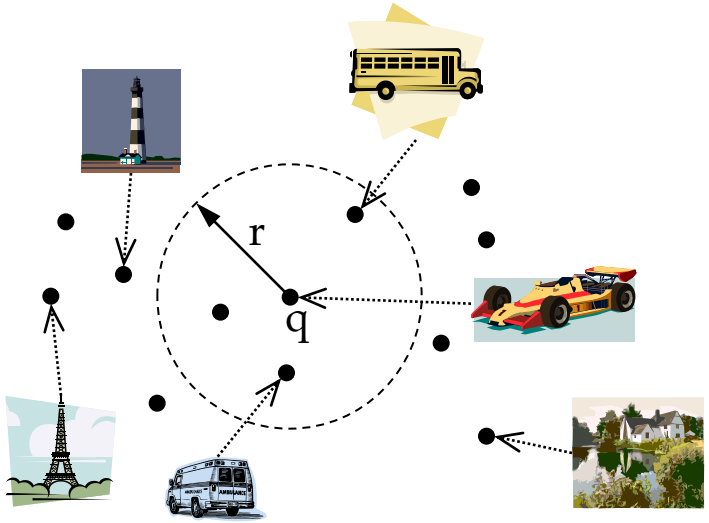
Motivation

- Our objective – to develop an engine for efficient search in unstructured P2P networks

- Problems:
  - Scalability – a large number of peers
  - Volatility – continual peers' churning
    ⇒ self-organizing systems
  - Similarity – complex data domains
    ⇒ metric space

Similarity Search: Metric Space

- Metric Space M is a pair $M=(D, d)$, where:
  - D is a domain of objects
  - d is a function measuring similarity between two objects
- Similarity range queries $R(q, r)$

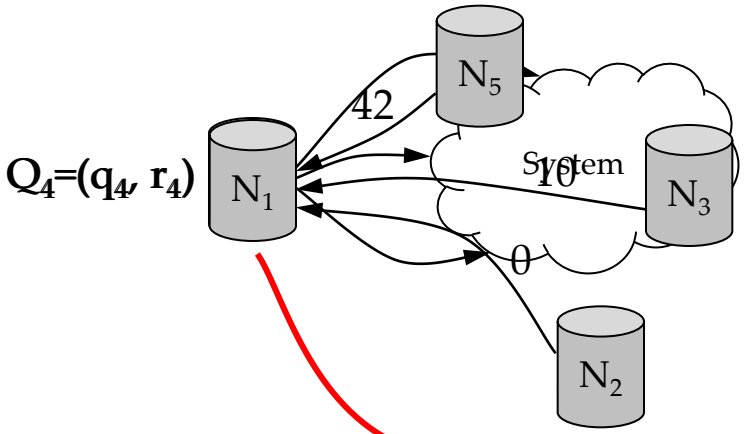


Metric Social Network

- Metric Social Network
  - A similarity search system for unstructured P2P networks
    - A set of peers interconnected by semantic links
    - Peers are independent and equal in functionality
    - There is no global control mechanism
    - Based on self-organizing principles:
      - Scalability
      - Adaptability
      - Robustness
    - Peer's schema:
      - Data repository (e.g., image features)
      - Routing table

Metric Social Network: Routing Table

- Routing table:
  - Exploration peers
  - Query history – based on answers to the processed query
    - Acquaintance – peer returning the largest part of the answer
    - Friends – peers returning a non-empty answer



Routing table

Exploration peers
N_8 N_{10} N_{13}

Query history

	Q	Friends	Acq.	Answer size	Confidence
E_1	q_1 r_1	N_7 N_2 N_4	N_7	189	0.82
E_2	q_2 r_2	N_6 N_7	N_6	13	0.95
E_3	q_3 r_3	N_1	N_1	7	0.30
E_4	q_4 r_4	N_5 N_3	N_5	52	0.74

Metric Social Network: Query Routing

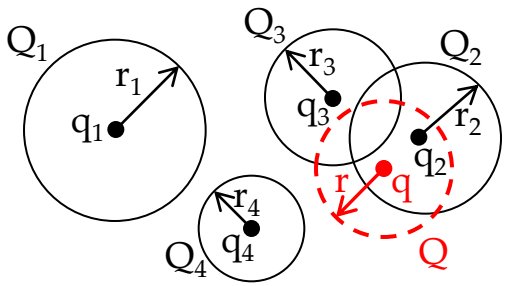
- At each peer, a query $Q(q, r)$ is processed as follows:
  - Take the most relevant entries E_i to Q
    - Exploitation – forward Q to the acquaintances of these entries
    - Exploration – forward Q to a certain number of exploration peers
  - Routing stops when no better acquaintance exists
    - Evaluate Q on the local data repository
    - Ask all friends of the most relevant entries to evaluate Q as well

Routing table

Exploration peers
N_8 N_{10} N_{13}

Query history

	Q	Friends	Acq.	Answer size	Confidence
E_1	q_1 r_1	N_7 N_2 N_4	N_7	189	0.82
E_2	q_2 r_2	N_6 N_7	N_6	13	0.95
E_3	q_3 r_3	N_1	N_1	7	0.30
E_4	q_4 r_4	N_5 N_3	N_5	52	0.74



Metric Social Network: Experimental Evaluation

- System size: 2,000 peers
- Data sets:
  - Synthetic – 100,000 2-d vectors
  - Real-life (CoPhIR image features) – 100,000 282-d vectors

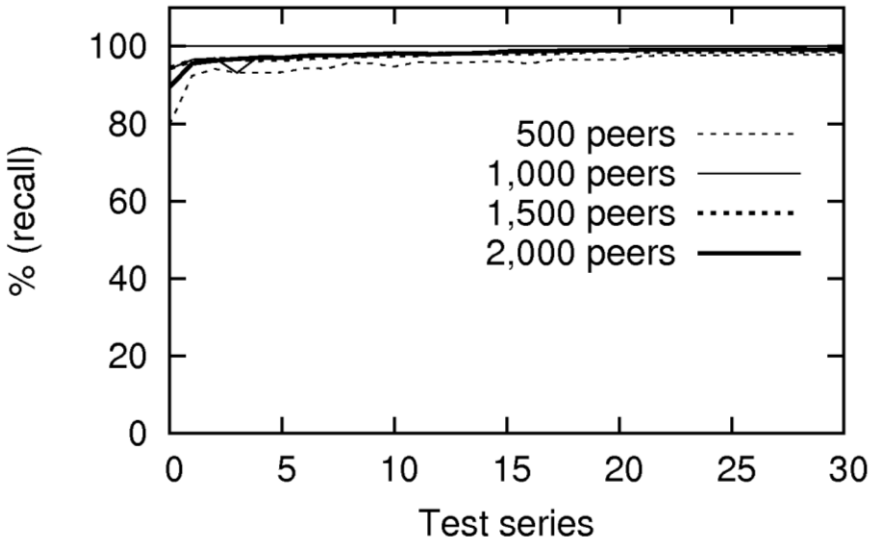
    $\Rightarrow$ each peer maintains 50 data objects
- Experimenting – repeating the batch of:
  - Training series – 50 queries executed at random peers
  - Test series – 5 queries executed at predefined peers

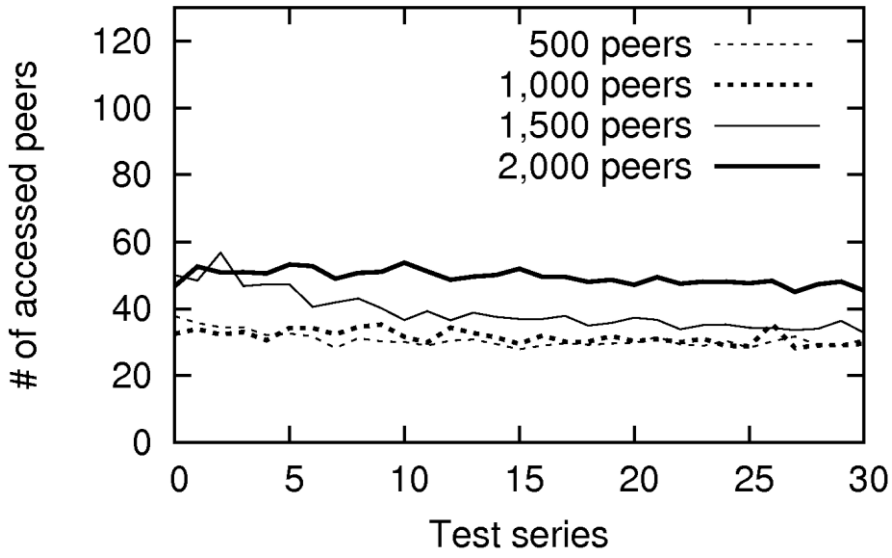
Metric Social Network: Experimental Evaluation (cont.)

- Measures:
  - Recall [%] – ratio between the size of the answer of our system and the size of the complete answer
  - Total costs [# of peers] – number of peers contacted by the routing algorithm in order to process a query
  - Optimal costs [# of peers] – number of peers in the system that contain data relevant to a query

Metric Social Network: Experimental Evaluation (cont.)

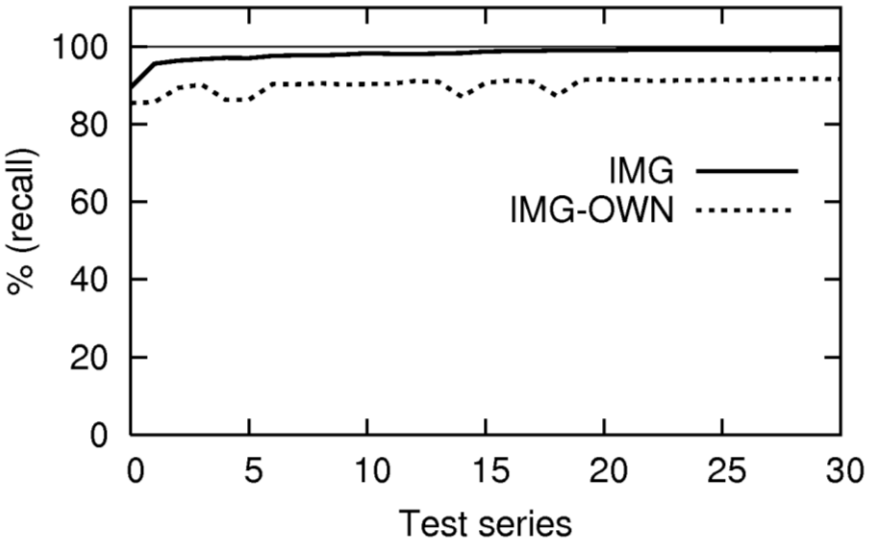
- Scalability evaluation (image features)
  - Very high recall – almost 100%
  - Low costs – 50 peers (out of 2,000) contacted on average

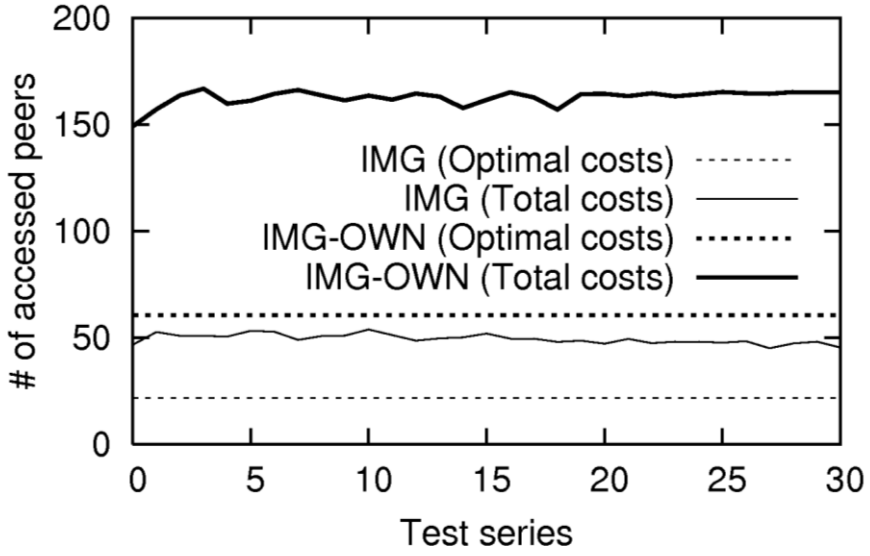




Metric Social Network: Experimental Evaluation (cont.)

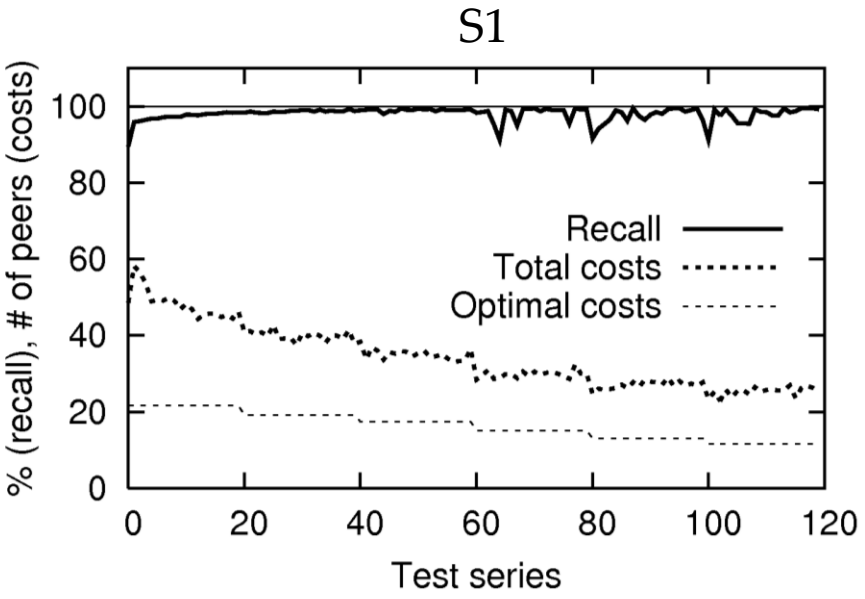
- Adaptability to data distributions (image features)
  - IMG – semi-clustering principle
  - IMG-OWN – random data distribution

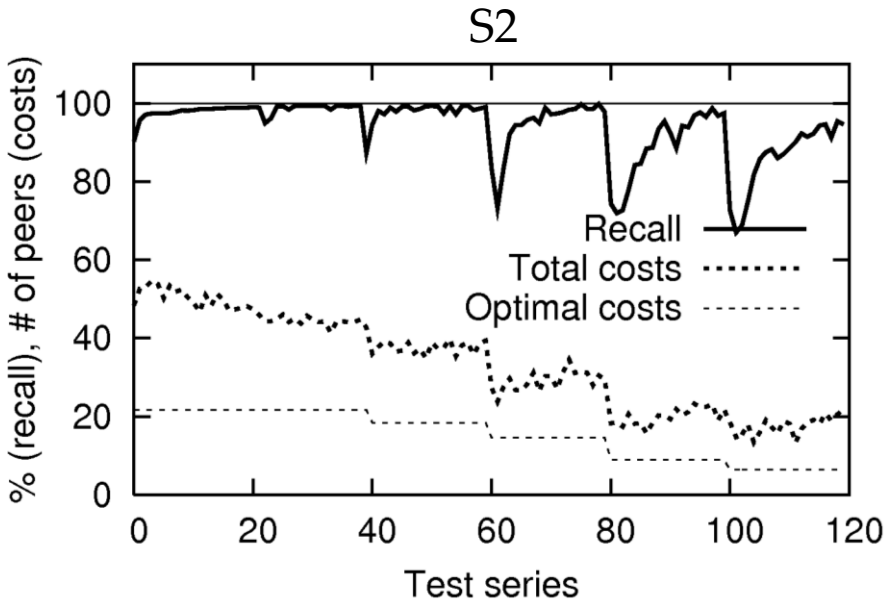




Metric Social Network: Experimental Evaluation (cont.)

- Resilience to disconnections of peers (image features)
  - After each 20th test series:
    - S1 – 200 random peers were disconnected
    - S2 – 200 the most knowledgeable peers were disconnected





Conclusions and Future Research Directions

- Main achievements:
  - Prototype of Metric Social Network
  - Experimental evaluation of scalability, adaptability, and robustness

- Future research directions:
  - Advanced experiments on peers' churning
  - New routing algorithms optimizing search costs

Questions?

Thank you for your attention.