A Self-organizing System for Large-scale Content-based Information Retrieval

Stanislav Bartoň, **Vlastislav Dohnal**, Jan Sedmidubský, Pavel Zezula

Faculty of Informatics, Masaryk University
Brno, Czech Republic

SemWeb Svratka
November 30–December 2, 2008

Outline

- Motivation
- Approaches
  - Metric Space
  - Self-organizing Systems
- Metric Social Network
  - Architecture
  - Query Routing
  - Experimental Trials
- Future Work

Motivation

- Our aim – to develop an engine suitable for searching in large networks

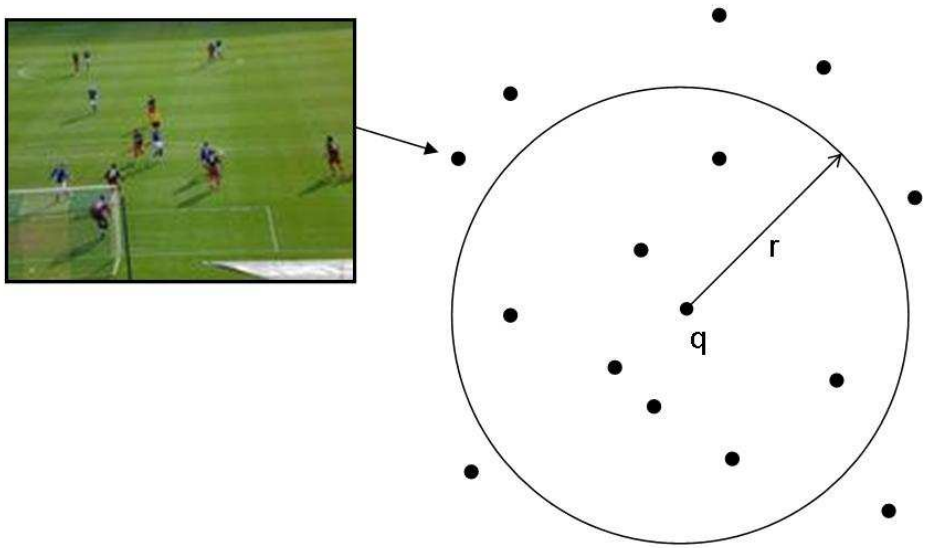
Motivation

- Our aim – to develop an engine suitable for searching in large networks

- Problems:
  1. Domains of complex objects
     - Non-sortable
     - Similarity-based (content-based) reasoning ⇒ Metric space
  2. Huge quantities of data
     - Exponential growth
     - Scalability problem ⇒ Self-organizing systems

Metric Space

- Metric space $\mathcal{M}$ is a pair $\mathcal{M} = (\mathcal{D}, d)$, where:
  - $\mathcal{D}$ is a set of objects – points in the metric space
  - d is a metric function measuring a distance (*similarity*) between two objects



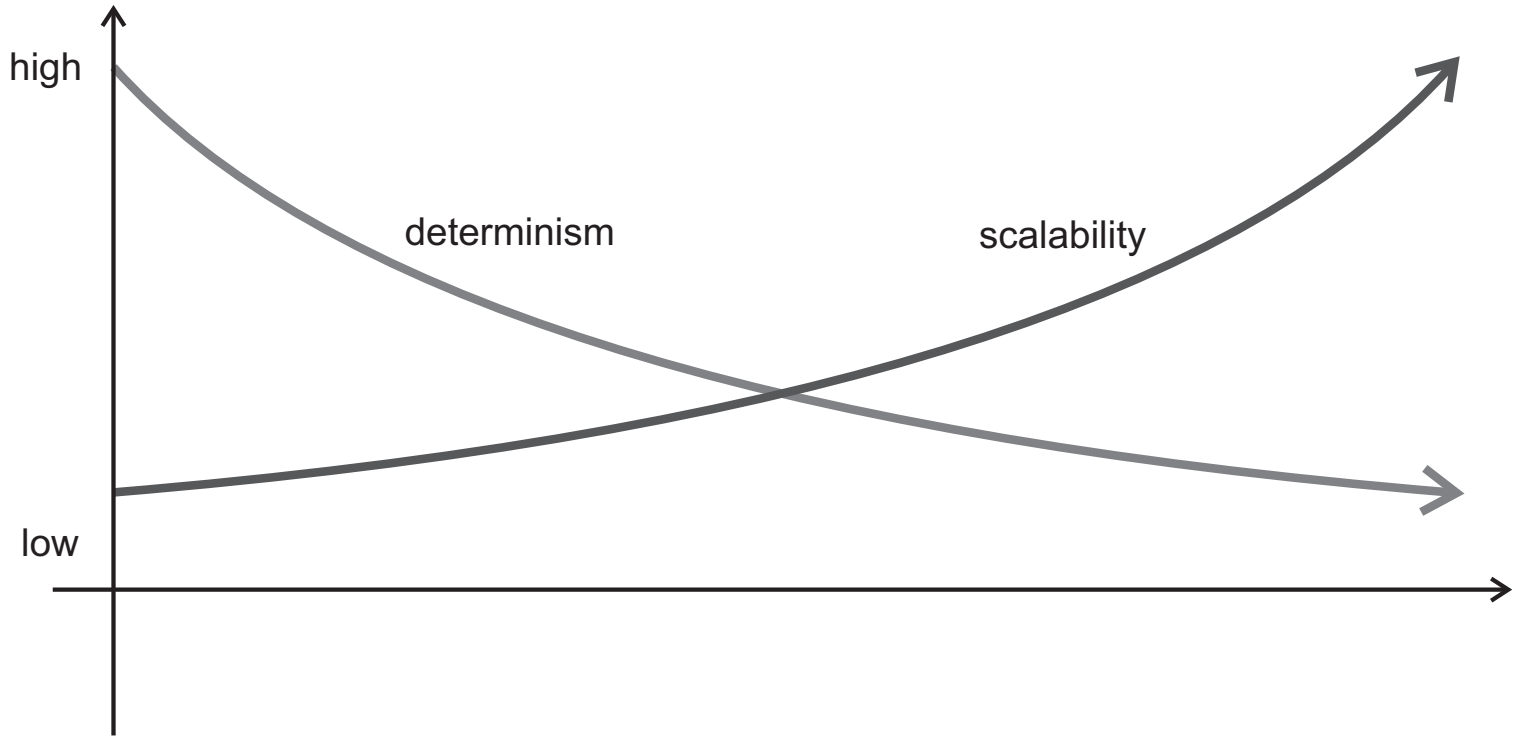
- Queries:
  - Range query $R(q, r)$
  - Nearest-neighbor query $NN(q)$

Search Problem

- Scalability – increasing amount of data, number of users (queries)

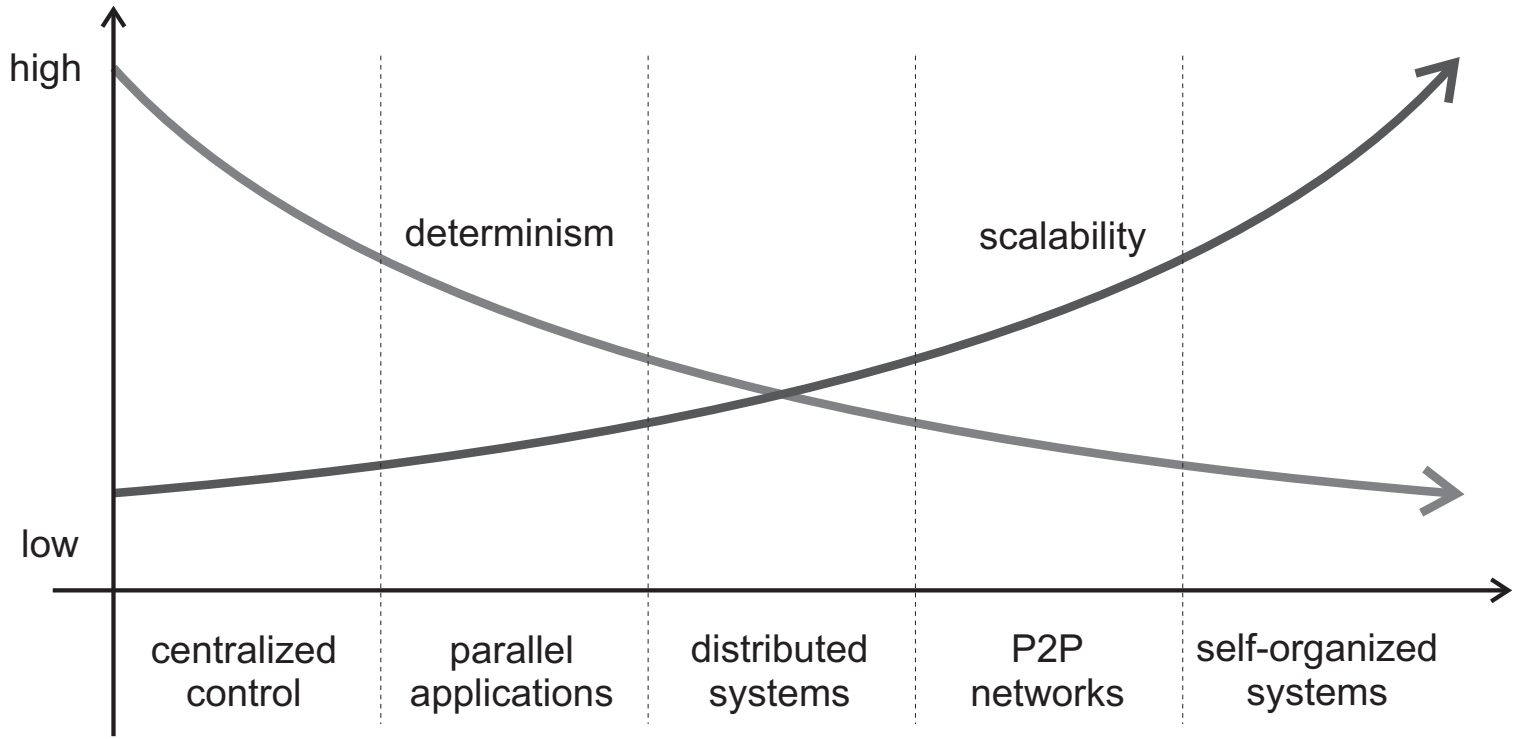
Search Problem

- Scalability – increasing amount of data, number of users (queries)
- Determinism:
  - exact match $\longrightarrow$ similarity
  - precise answer $\longrightarrow$ approximate answer



Search Problem

- Scalability – increasing amount of data, number of users (queries)
- Determinism:
  - exact match $\longrightarrow$ similarity
  - precise answer $\longrightarrow$ approximate answer



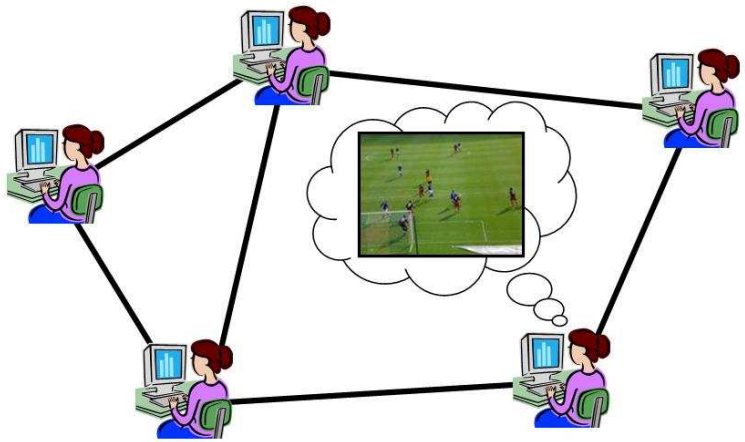
Self-organizing Systems

- A set of interacting components creating a desired outcome
  - Evolves in time and space
  - Inspired in biology, sociology, physics, ...

- Properties:
  - Scalability
  - Adaptability
  - Robustness

Self-organizing Search Systems

- Our aim – apply principles of self-organization to build a robust search engine
  - A desired outcome = a search engine



- Properties:
  - Scalability
  - Adaptability
  - Robustness

Metric Social Network

Metric Social Network (mSN) =

= Metric space + Self-organization principles

Metric Social Network

Metric Social Network (mSN) =

= Metric space + Self-organization principles

- Self-organizing network for similarity searching
- Supports range queries (answers are approximate)

- Structure:
  - Peers – computers
  - Relationships – logical connections between peers

relationships $\neq$ physical connections between peers

Structure

Peers

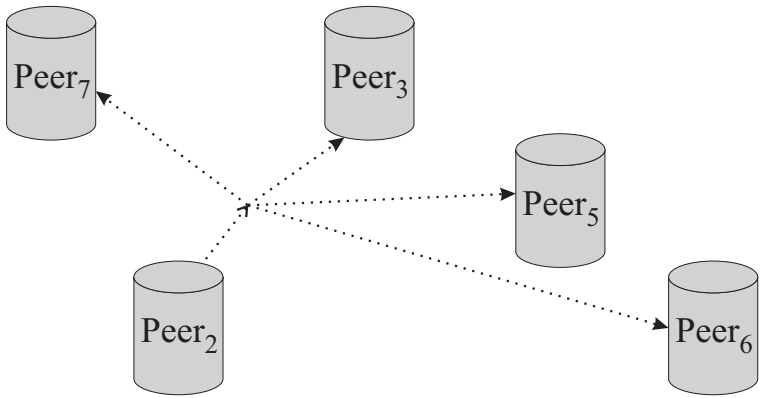
- Data – local data governed by the peer, e.g. images
- List of random peers
- Query history – experience with previous querying

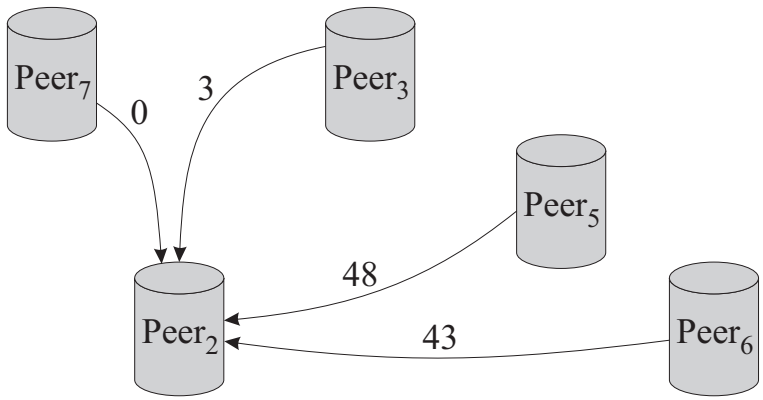
Relationships

- Exploited by the query-routing algorithm
- Based on the social-network paradigm:
  - Acquaintance relationships – navigation purposes
  - Friend relationships – identify peers similar in content

Relationships

- Created according to peers' answers to the processed query

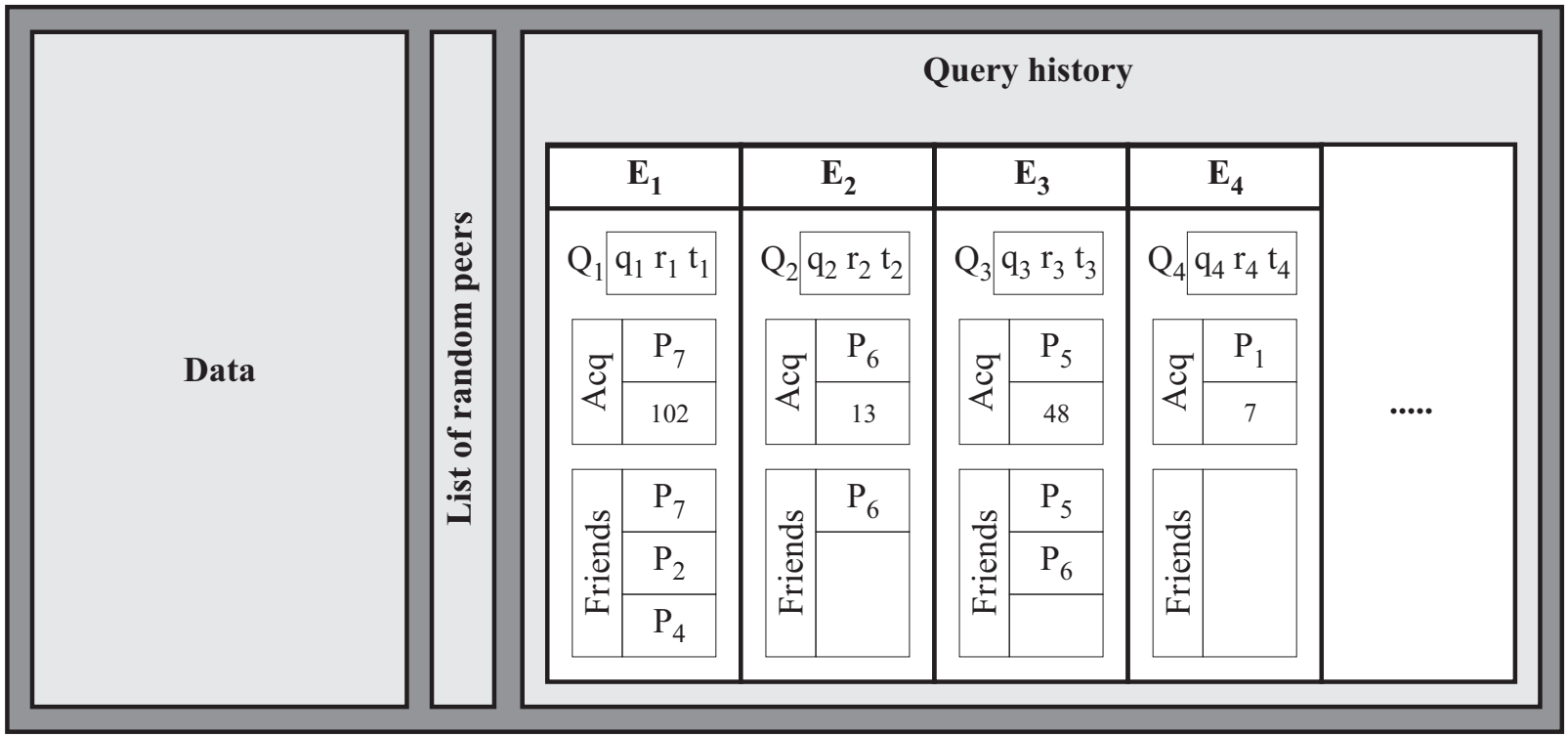




- Acquaintance – peer with the best quality of the answer
  - $Acquaintance(Q) = Peer_5$
  - Acquaintance relationship: between $Peer_2$ and $Peer_5$
- Friends – peers with the significant quality of the answer
  - $Friends(Q) = \{Peer_5, Peer_6\}$
  - Friend relationships: between each of two friends

Peer Anatomy

- Query history – a list of *entries* $E_1, \ldots, E_n$ containing metadata about queries processed so far
  - Query identification (query object, radius, timestamp)
  - Acquaintance
  - List of friends



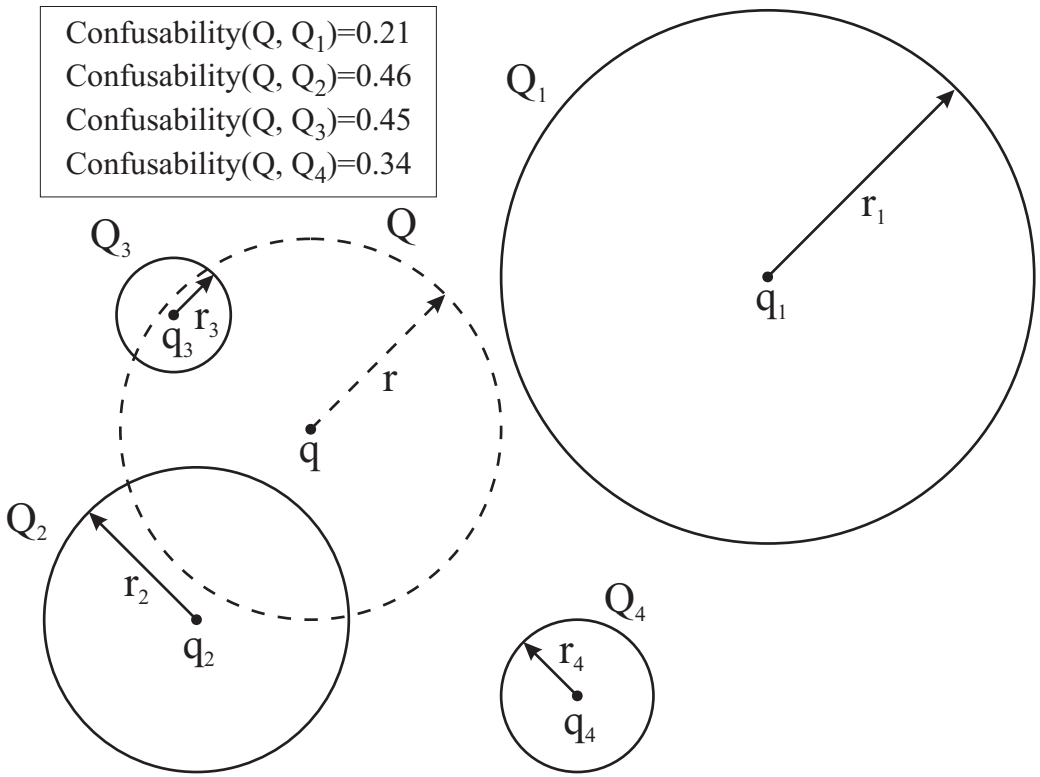
Searching

- User poses a query at the query-issuing peer
- Each peer selects the peers to which the query is forwarded
  - Exploitation strategy
    - Take the most relevant peers to the query
  - Exploration strategy
    - Take some random peers
- The query is evaluated on local data of contacted peers
- The quality of answers is determined by the query-issuing peer
- The query history is updated

Confusability

- Confusability expresses the similarity between two range queries Q_1 and Q_2 with timestamps

$$Confusability(Q_1, Q_2) = w_D \cdot D(Q_1, Q_2) +$$
$$+ w_I \cdot I(Q_1, Q_2) + w_T \cdot T(Q_1, Q_2)$$



Query Routing

1. Retrieve five most confusable entries from the query history
2. Determine *max_confusability* of these five entries
3. Route the query
   - Exploitation strategy
     - Depending on *max_confusability*, determine the number n of entries to use

max_confusability $\geq$	0.90	0.65	0.40	0.15	0.00
n	1	2	3	4	5

     - Get acquaintances from these entries and forward the query
   - Exploration strategy
     - With a probability $1 -$ *max_confusability*, pick a peer from the list of random peers and forward the query to this peer.

Query Routing (cont.)

- Query forwarding stops when:
  - Maximum hop count is reached
  - An acquaintance of higher quality does not exist
- The query is evaluated on local data of:
  - The most relevant acquaintances
  - Friends of the most relevant acquaintances respecting the query
- The answers are returned to the query-issuing peer
  - The qualities of answers are computed
  - New relationships are established

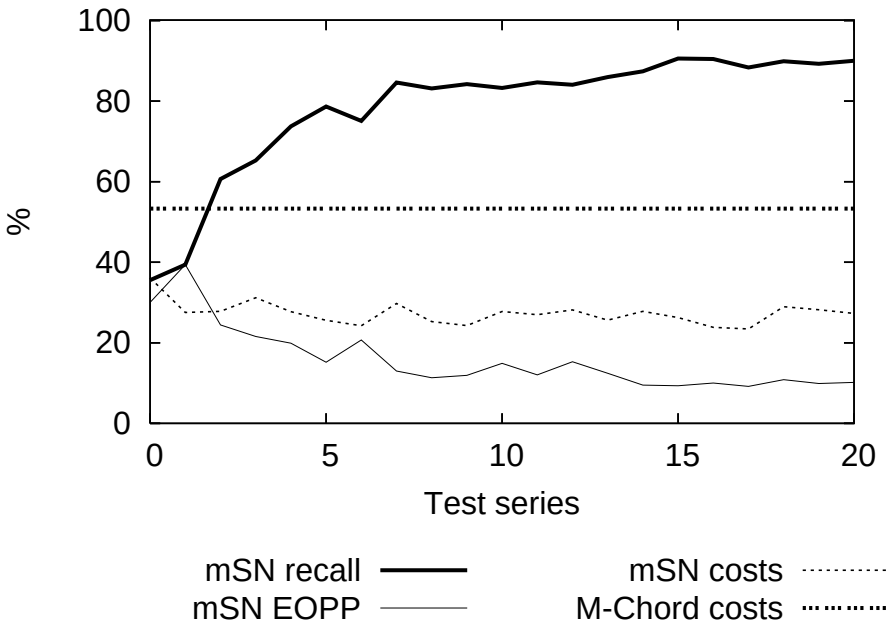
Experimental Trials

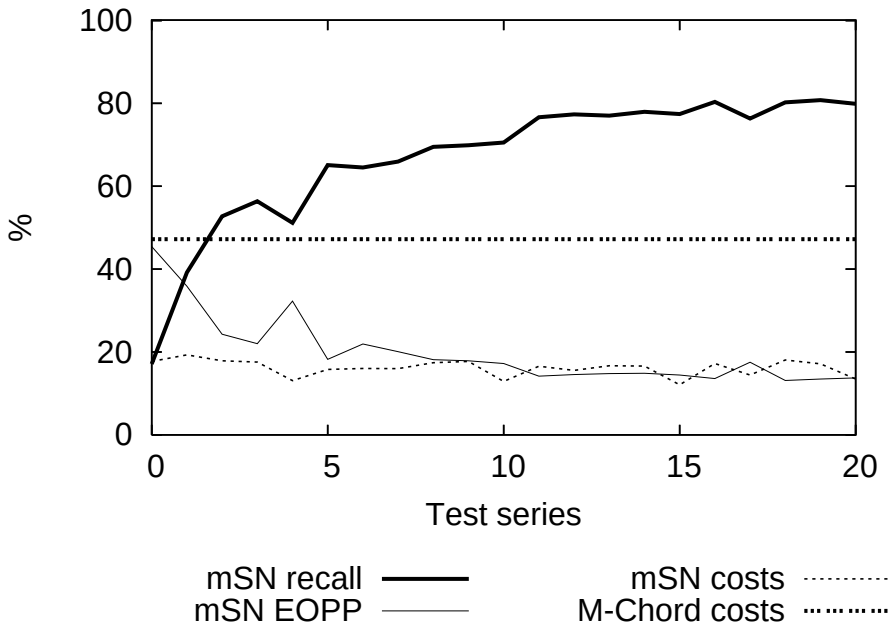
- Networks
  1. 500 peers indexing 2,500,000 images
  2. 2,000 peers indexing 10,000,000 images

- Measures
  - Costs – a ratio between the number of accessed peers and the number of all peers in the network in percents
  - Recall – a ratio between the sizes of mSN answer and the precise answer in percents
  - EOPP – a normalized error on peers' positions expressing the inaccuracy of approximate answer of mSN

Experimental Trials (cont.)

- A batch of random 50 range queries between each of two test series
- Each test series consisted of fixed 20 range queries
- Results compared to *M-Chord* – a structured P2P network





Future Work

- Dynamicity
  - Massive peers' churning
  - Joining two networks
- Knowledge management
  - Positive / negative feedback of querying

Questions?

Thank you for your attention.

- Supported by:
  - National research project 1ET100300419
  - EU IST FP6 project 045128 – SAPIR
  - Czech Science Foundation projects 201/07/P240 and 102/05/H050