

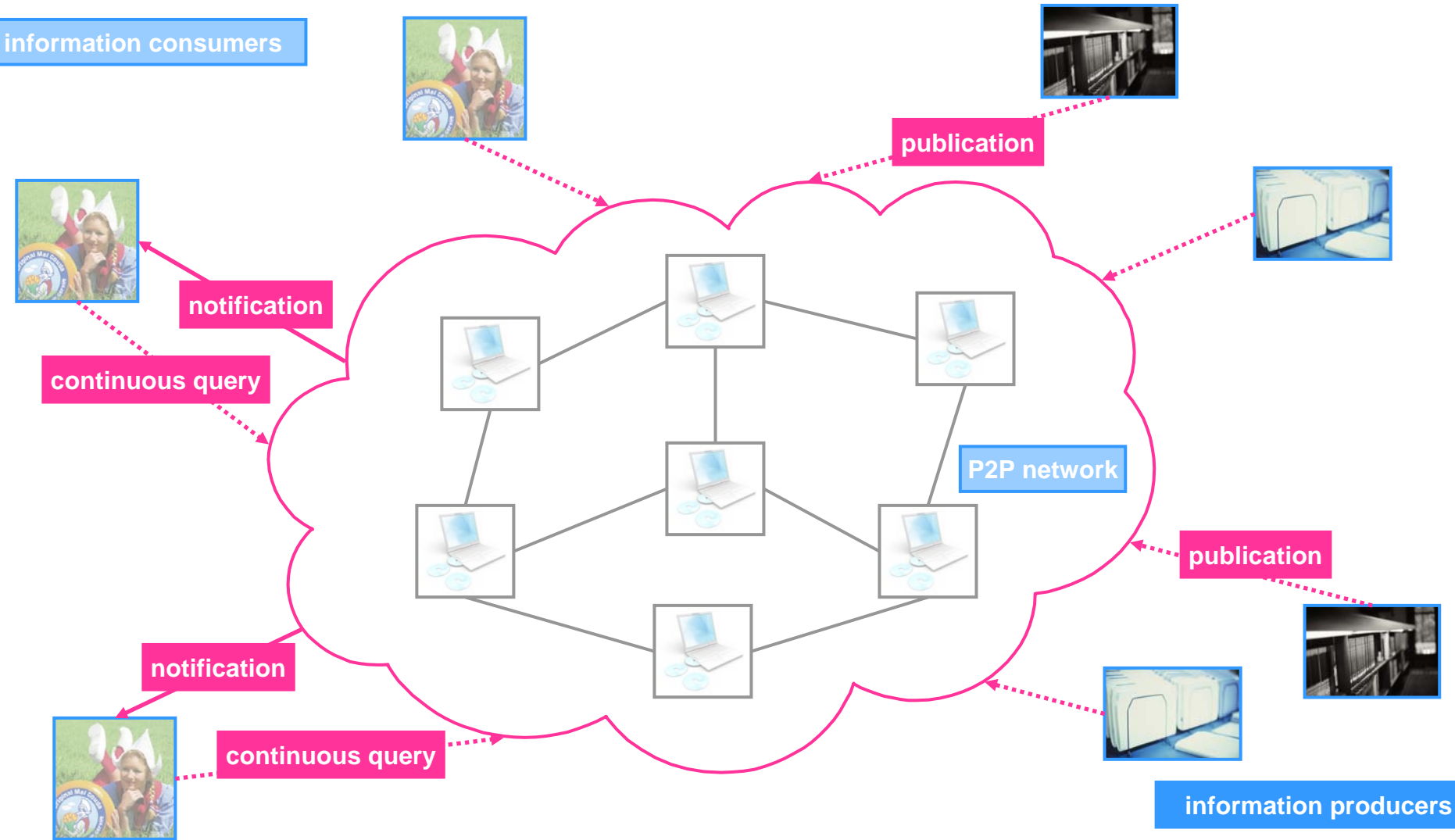


Node Behavior Prediction for Large-Scale Approximate Information Filtering

Christian Zimmer, Christos Tryfonopoulos, Klaus Berberich, Gerhard Weikum
Max-Planck Institute for Informatics, Saarbrücken, Germany

Manolis Koubarakis
National and Kapodistrian University of Athens, Greece

Motivation



Motivation



Peer-to-Peer

- **Autonomous peers** with local data collections.
- Structured, unstructured, and hierarchical networks.
- Peers have **different interests** on publishing and subscribing.

Information Filtering

- **Subscriptions** (continuous queries), **publications**, and **notifications**.
- Mainly centralized approaches (InRoute, SIFT, DIAS).
- IR & IF: "**The two sides of the coin**" [Belkin & Croft, CACM, 1992].

We combine it!

- **P2P-IF**

Outline of the Talk



- Motivation
- Introduction to P2P-IF
- MAPS Architecture
- Problem Statement: Node Behavior Prediction
- Double Exponential Smoothing
- Selective Method Approach
- Experimental Evaluation
- Conclusions & Open Questions

Introduction to P2P-IF



Information Filtering

- aka **Continuous Querying / Publish/Subscribe**
- User (node) **subscribes** a **continuous query** (or **subscription**) to the P2P network and get **notified** when documents of interest are **published**.

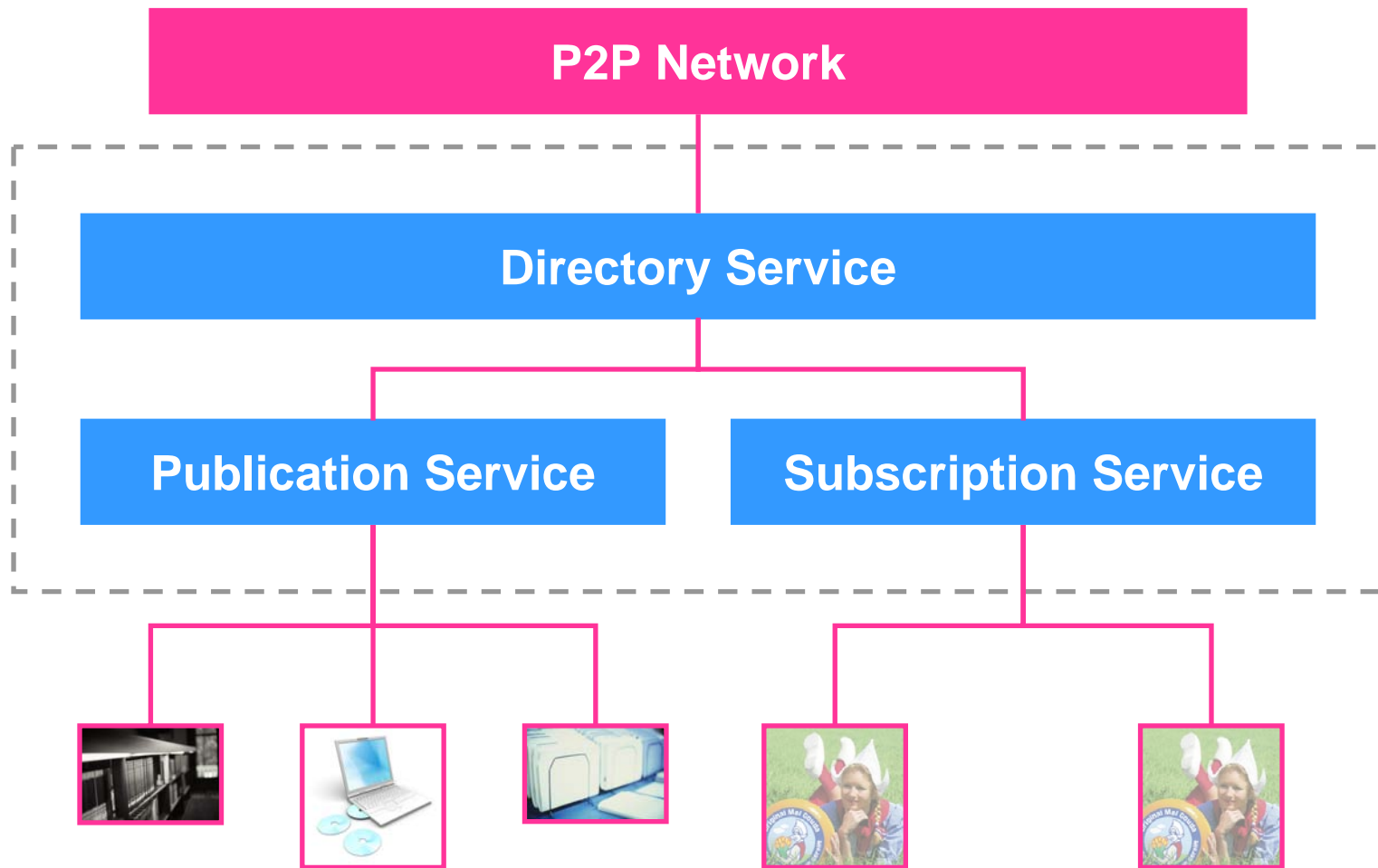
Exact Information Filtering

- Existing Pub/Sub approaches with **underlying hypothesis**: subscribers are interested in notifications for all published documents.
 - At the first glance: **That's what we want!**
 - What about **scalability**? Why **notifications for all** documents?

Approximate Information Filtering

- Our Approach: **MAPS (Minerva Approximate Publish/Subscribe)**
 - Do not try to get all notifications. **Win scalability and efficiency**.
 - Subscribers only **monitor** most promising nodes.
 - **Node Selection** is critical decision.

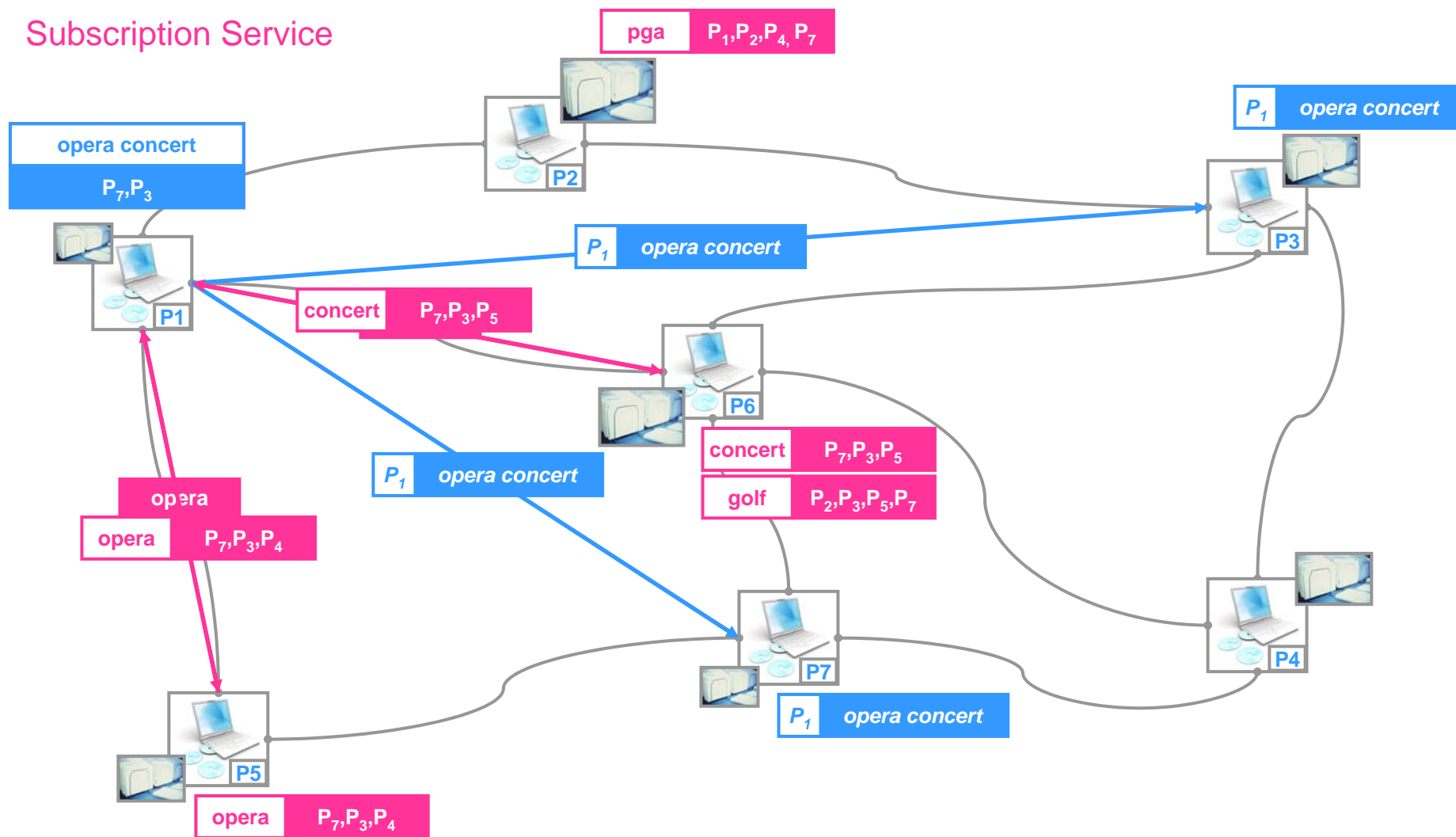
MAPS Architecture



MAPS Architecture



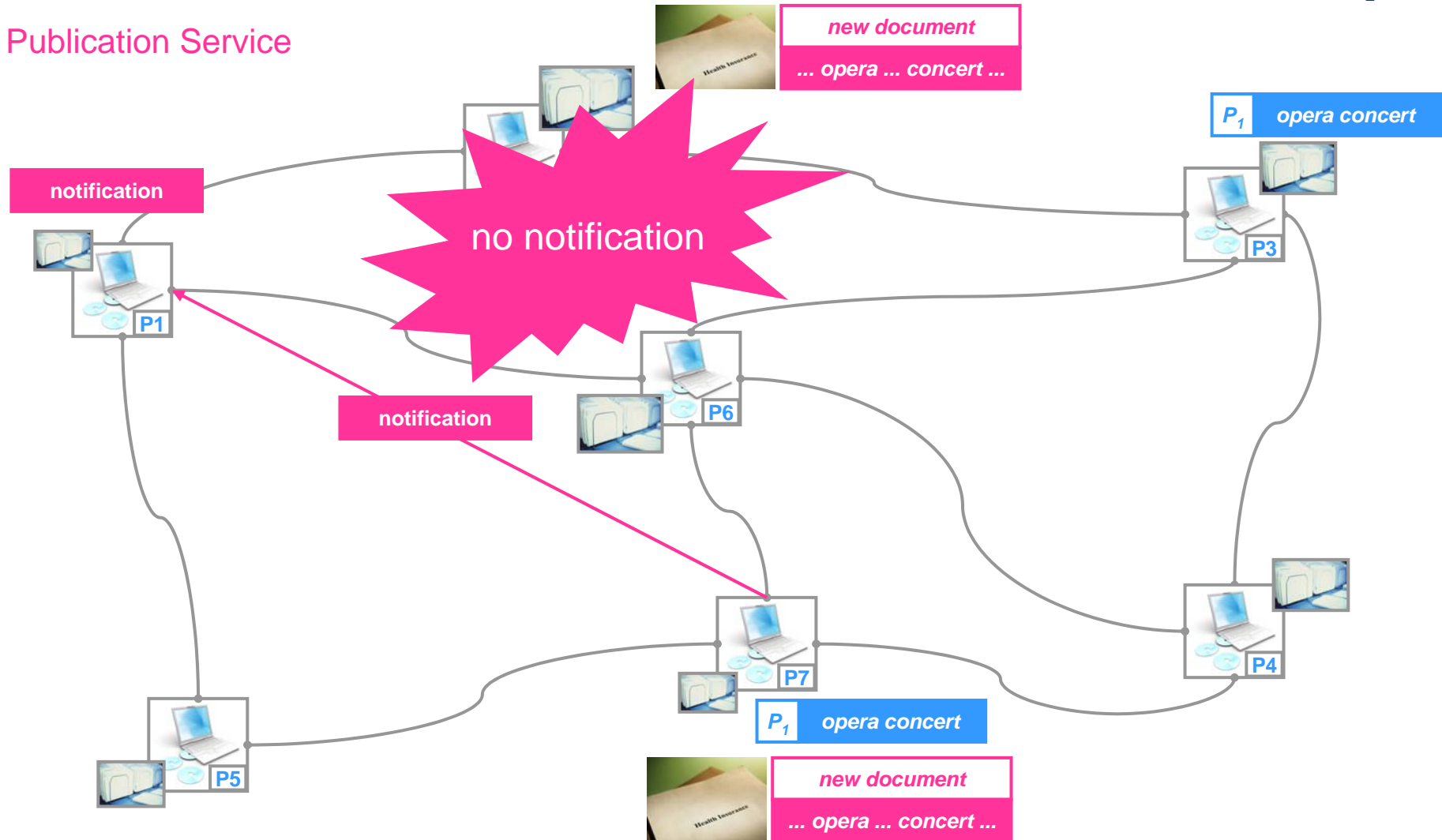
Subscription Service



MAPS Architecture



Publication Service



Problem Statement: Node Selection



- **Question:** Which nodes in the network are promising candidates to satisfy the continuous query with appropriate documents published in the future?
- Subscriber node uses **directory statistics** to compute **publisher scores** and to **rank publisher nodes**.
- Subscriber node forwards the continuous query to the **top-ranked publisher nodes (monitoring)**.
- Continuous queries need to get **periodically re-positioned**.



Problem Statement: Node Selection



Resource Selection

- Score to identify **authorities** (nodes with appropriate local document collections)
- **Resource Selection algorithms**: CORI, GIOSS, tf-idf-based approaches (well-known techniques in Information Retrieval field)
- Today, we **do not focus on Resource Selection** (although it can improve node selection).

Node Behavior Prediction

- Score to represent the **likelihood to publish relevant documents in the future**.
- MAPS **predicts the number of relevant documents** per publisher node in the next time-period based on **time-series analysis of document frequencies**.
- Our prediction technique uses **double exponential smoothing** to determine $\delta df_{p,t}^*$ as the number of relevant documents node **p** will publish concerning term **t** in the next time period.
- $\text{pred}(p,q) = \sum_{t \in q} \log(\delta df_{p,t}^* + 1)$
- Today, we **focus on analysing Node Behavior Prediction** (including double exponential smoothing)

Double Exponential Smoothing



Time-Series Analysis

- Techniques from **time-series analysis** predict **future values** based on past observations (observed values).
 - time-series values x_1, \dots, x_{n-1}
 - predicted value x_n^*
- All **techniques differ** in their assumptions (e.g., **moving average techniques** or **exponential smoothing techniques**).

Exponential Smoothing

- **recent values more important than old values.**
 - **Single** Exponential Smoothing: not trends, no seasonality
 - **Double** Exponential Smoothing: trends, but no seasonality
 - **Triple** Exponential Smoothing: trends and seasonality
- MAPS uses **Double Exponential Smoothing**: **interested in trends**, but seasonality only necessary for extrem long-lasting queries.
 - $x_n^* = L_n + T_n$
 - $L_n = \eta \cdot x_{n-1} + (1 - \eta) \cdot (L_{n-1} + T_{n-1})$ [Level]
 - $T_n = \gamma \cdot (L_n - L_{n-1}) + (1 - \gamma) \cdot T_{n-1}$ [Trend]

Double Exponential Smoothing



Investigation of Double Exponential Smoothing

- Influence of η and γ by looking at **all possible combinations** from 0 to 1 in steps of 0.1 (121 parameter combination).
- Eight different series of values (behaviors):
LOG_INC, LOG_DEC, LIN_INC, LIN_DEC, QUAD_INC, QUAD_DEC, EXP_INC, EXP_DEC.
- Comparing the **prediction errors** depending on η and γ
 - **Result 1:** high variation of prediction errors depending on the choice of the parameter combination.
appropriate parameter selection necessary
 - **Result 2:** no single parameter combination yields to lowest prediction errors for all behaviors. The best combination for one behavior does not necessarily result in satisfying predictions for another one.
global parameters for all behaviors not ideal

Overall Conclusion

- The choice of the parameters η and γ needs to be **adapted to the observed value behavior.**

Selective Method Approach



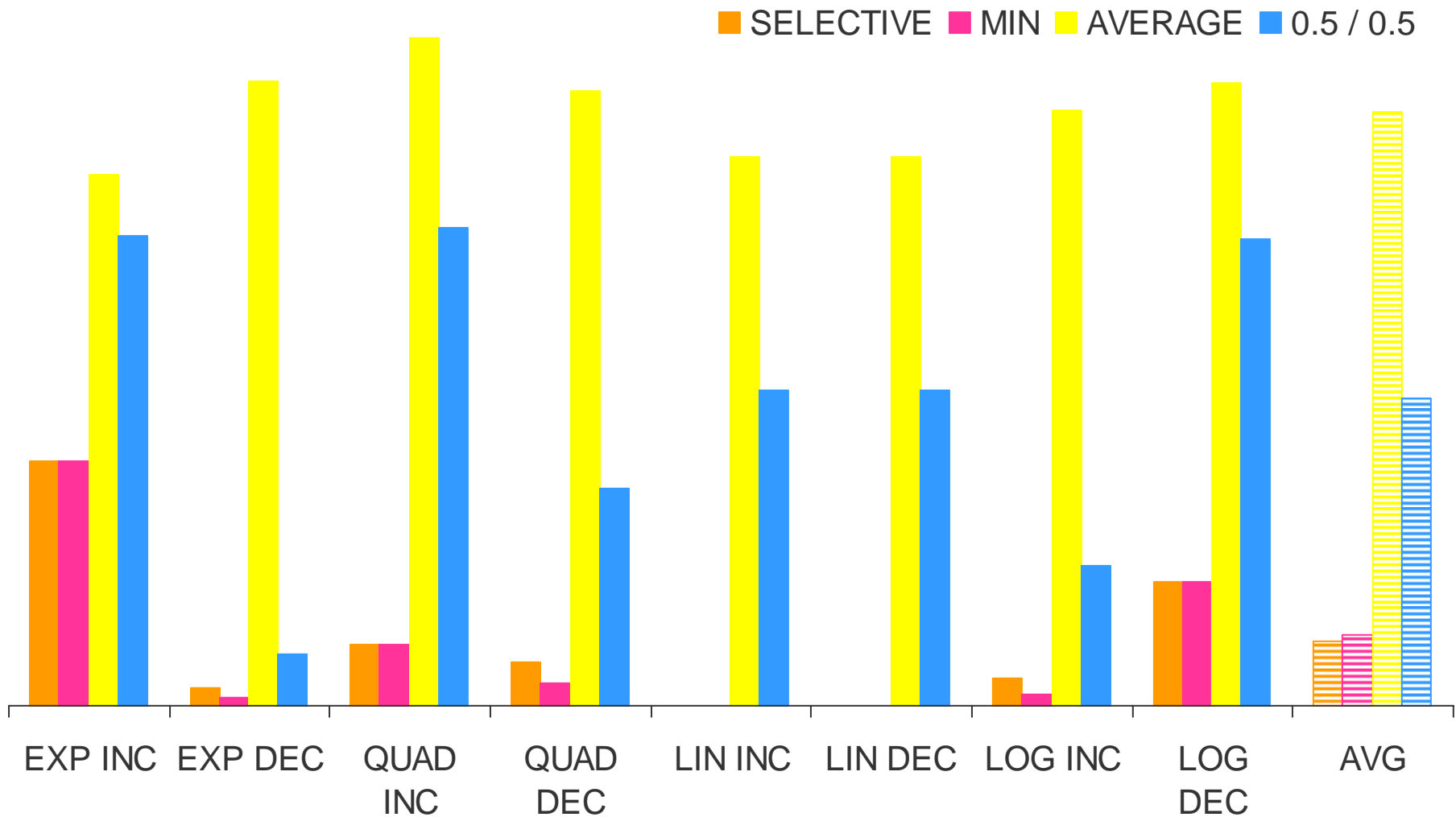
- ... to adapt the parameter setting to the given scenario.

- **Basic Algorithm of Selective Method**
 1. Consider x_1, \dots, x_{n-1} as observed time-series values.
 2. Values x_1, \dots, x_{n-2} are used to predict the already observed value x_{n-1} .
 3. Let $x_{n-1, \eta, \gamma}^*$ denote this predicted value for all parameter combinations of η and γ .
 4. Next, select the parameter setting with smallest error concerning the real observed value x_{n-1} .

- **Selective Method** means, we always select the **most appropriate parameter setting concerning the last observed value**.

- Obviously, **four observed values** are indispensable to properly predict the last observed value (although two values are sufficient to apply the algorithm).

Selective Method Approach



Selective Method Approach



How to use it in MAPS

- Given a **continuous query** and the collected **publisher statistics** out of the directory.
- Subscriber node applies the **Selective Method approach** to the observed statistics of **each publisher node** to predict the number of documents in the next time-period.
- Subscriber uses **different parameter combinations** for all publishers depending on their publishing behavior concerning the continuous query.

Advantages

- **Only local computations** are needed (no additional communication between the nodes).
- Subscriber node does **not need to preselect an appropriate parameter combination** that can result in completely bad predictions.

Experimental Evaluation



Setup

- After the DHT is initiated, subscriber node collects in **four bootstrapping rounds** directory statistics to the queries.
- During the **six monitoring rounds**, the subscriber issues the continuous queries to the selected publisher nodes.
- All (publisher) nodes **publish documents during both phases** (depends on the experimental series).
- Before each monitoring round, the **subscriber node ranks publisher nodes** and **(re-)positiones the queries**.

Measurement

- **Recall**: ratio of total number of received notifications to total number of published and relevant documents (**average recall** over all (monitoring) rounds).
 - MAPS **Selective Method** approach
 - **Min / Max** recall by using global parameter setting
 - **Oracle** recall with always predicting the accurate statistics

Experimental Evaluation



Data Collection

- Document collection of ~ 2 million documents (focused Web crawl). Each document categorised to one of ten categories (e.g., Travel, Sports, or Finance).
- Size of categories: 68,000 to 325,000 documents.
- More than 500,000 different terms (without stopwords).

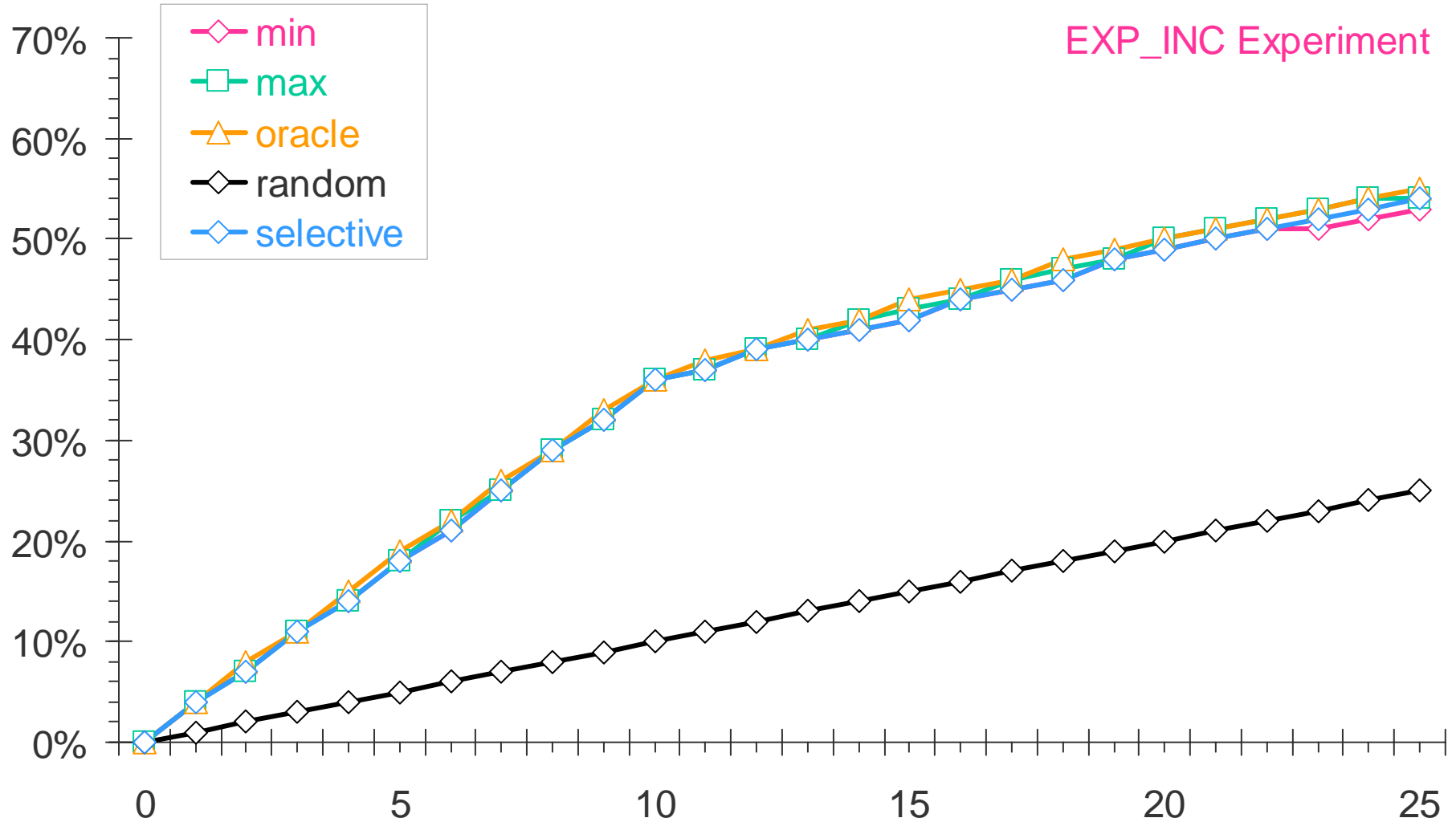
Continuous Queries

- Seven strong representative single-term queries: music, arts, sports, travel, hotel, offer, city.
- Advantage of single-term queries: direct dependency between correctly predicting values and recall.

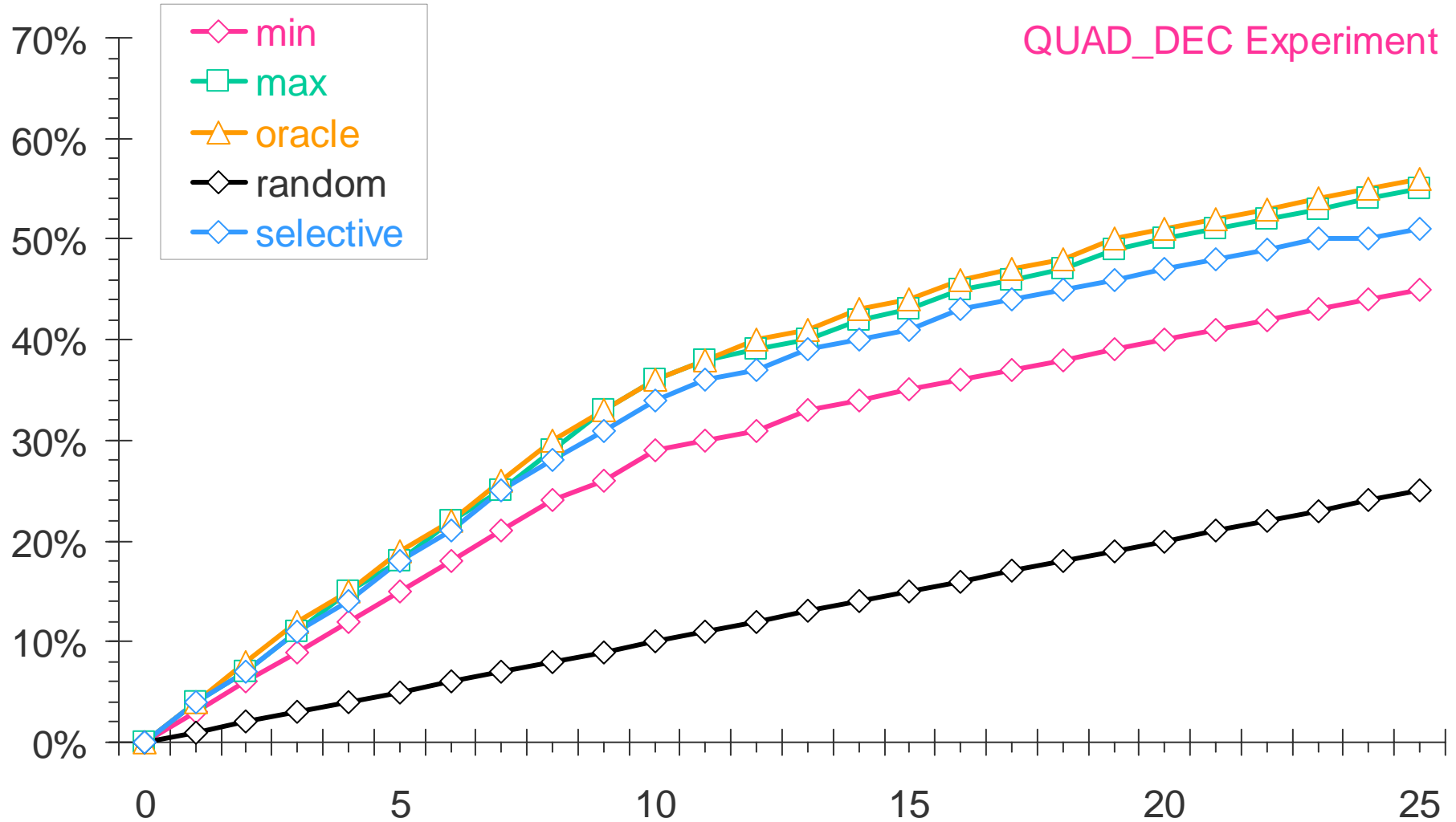
Network / Nodes

- 100 nodes / 10 nodes per category.
- Different publishing behaviors (including constant publishing of 300 documents) over the ten rounds of 0 to 600 documents per round.

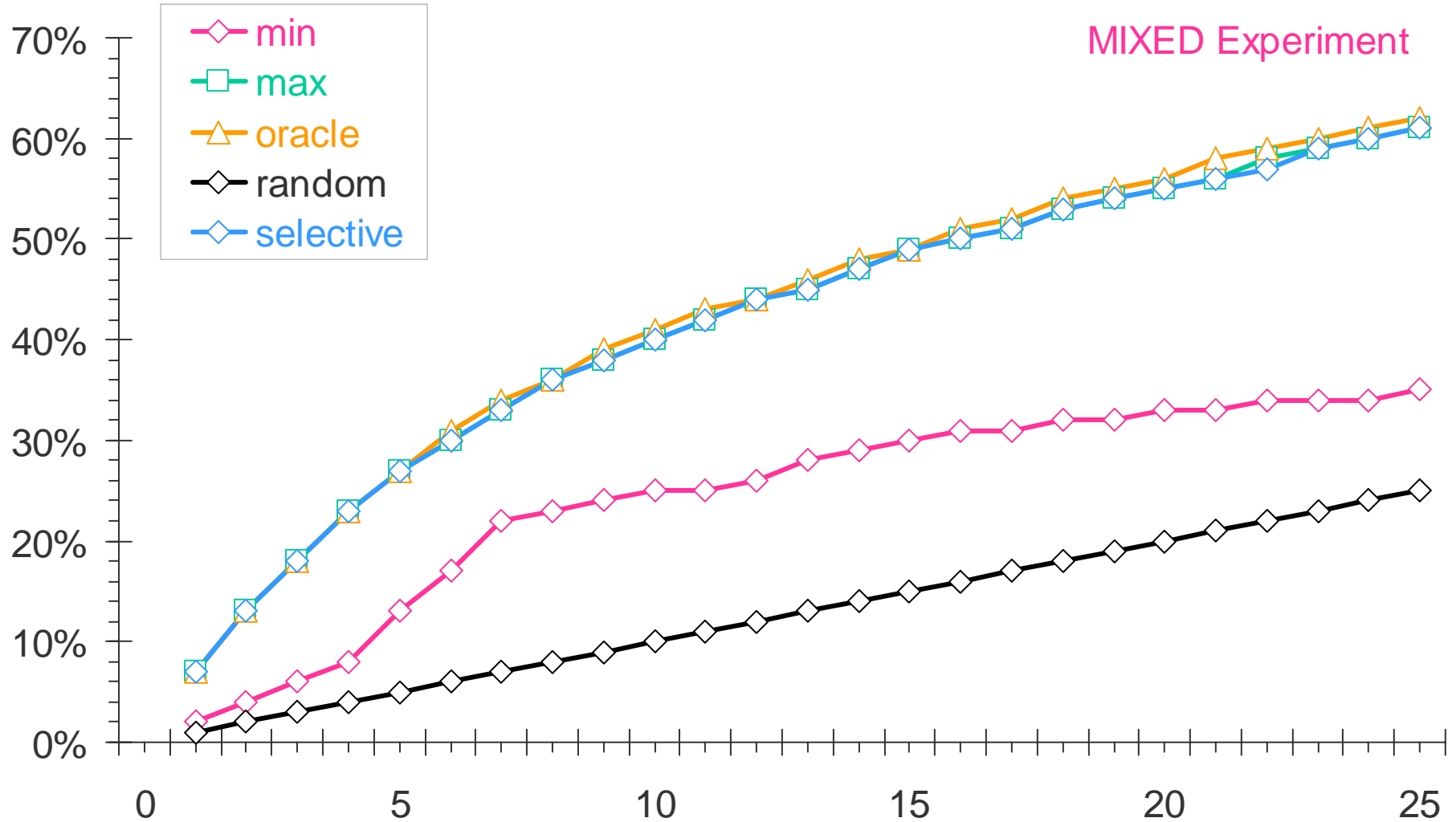
Experimental Evaluation



Experimental Evaluation



Experimental Evaluation



Experimental Evaluation



Summing-up of the Experimental Results

- Local auto-adjustment of prediction parameters is possible and results in **recall of 60%** by monitoring only **20% of the publisher nodes**.
- All approaches **outperform the random node selection** strategy.
- Recall level can **even be better than the best possible recall** using a global set of parameters (preselection).
- Selective Method **almost reaches recall level of an Oracle predictor**.
- **No additional communication costs are needed**.

Prediction Errors

- How exact can we predict the number of published and relevant documents in the next time-period.
- Selective Method approach **reduces errors of predicting the expected statistics** concerning the next time-period.

Conclusions & Open Questions



Conclusions

- MAPS: novel approach for approximate P2P-IF
- Subscriber nodes monitor the most promising publisher nodes for appropriate notifications matching a continuous query.
- Node Selection as critical decision is based on node behavior prediction using time-series analysis with double exponential smoothing.
- MAPS Selective Method approach adapts smoothing parameters to recognize publishing behavior of each publisher node concerning a continuous query without additional communication.

Open Questions

- Considering correlated termsets instead of single terms.
- Exact message costs analysis including directory message costs (in comparison to exact P2P-IF approaches).
- Integration of Information Retrieval and Information Filtering (one system, common message costs, and two functionalities).

Thank You For Your Attention!

Questions or Comments?