Building Web Query Subsumption Hierarchies

By Alexander Prohaska

Advisors: Georgiana Ifrim
Christos Tryfonopoulos

Challenges of Web Queries

- Average query length of 2.2 words
- Popular queries have length of 1.7 words
- Queries are ambiguous
- Huge vocabulary
- Constantly changing

Query Classification

- Defined by the „similar-to“ relation
  - contains two equivalent queries
  - for example „sport“ and „basketball“ are similar to each other
- Narrows the field of relevant documents
- Route queries to domain-specific databases

Query Subsumption

- Defined by the „is-a“ relation
  - generalization / specialization
  - for example „sport“ subsumes „basketball“ if „basketball“ is a specialization of „sport“ and „sport“ a generalization of „basketball“
- Arranges queries similar to Web directories
- Construct sets of relevant documents by combining sets from more special queries

Approach Restrictions

- No use of online resources, e.g. results from search engines

  $\Rightarrow$ independent offline approach
- No click-through data from search engines

  $\Rightarrow$ no proprietary and privacy issues
- Queries do not interfere with each other and are mapped once on their arrival

  $\Rightarrow$ no rebuilding of query subsumptions

$\Rightarrow$ usability in real applications

Overview

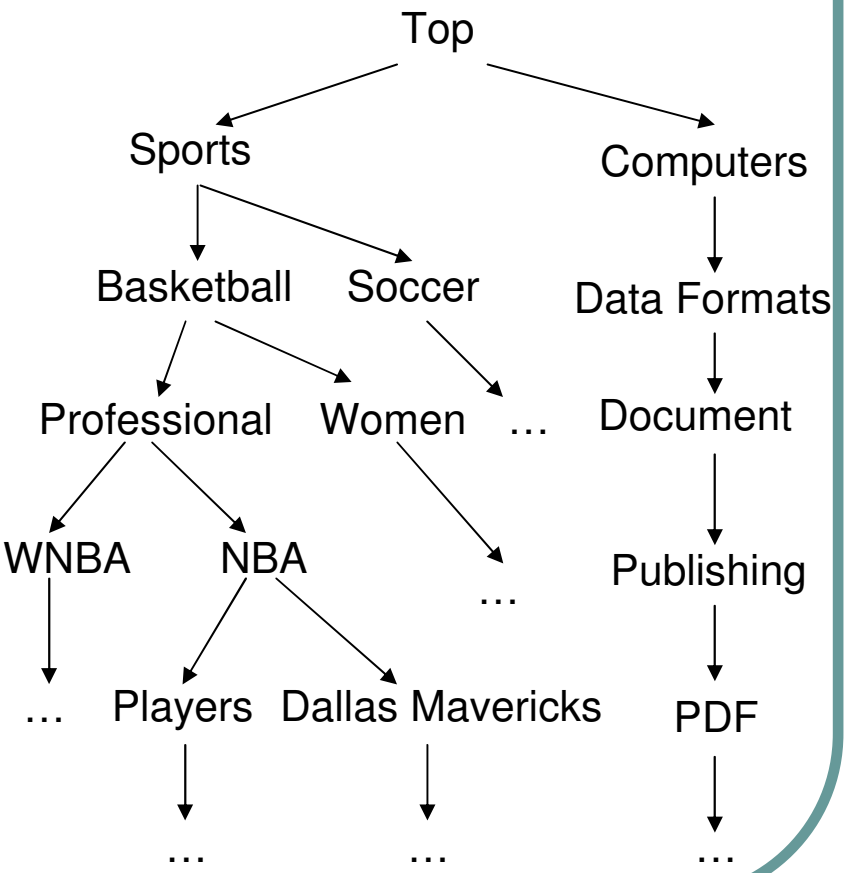
(1) Introduction
(2) Query-Driven K-Nearest Neighbors
(3) Evaluation Results
(4) Application

Query-Driven K-Nearest Neighbors

- Use knowledge base of Web directory
- Expand queries in kNN manner
- Map queries to Web directory structure
- Retrieve subsumption pairs and paths
- Build subsumption hierarchy

DMOZ Hierarchy

- More than 590.000 topics
- Up to 12 levels
- Over 4.5 million Web sites
- 15 top categories, e.g. Arts, Computers, Games, Health, Home, News, Recreation, Science, Shopping, Society, Sports



DMOZ Web Site Examples

Sports/Basketball/Professional/NBA/Players/J/Jordan,_Michael/

Link: http://digilander.libero.it/airmj/

Title: AirMJ

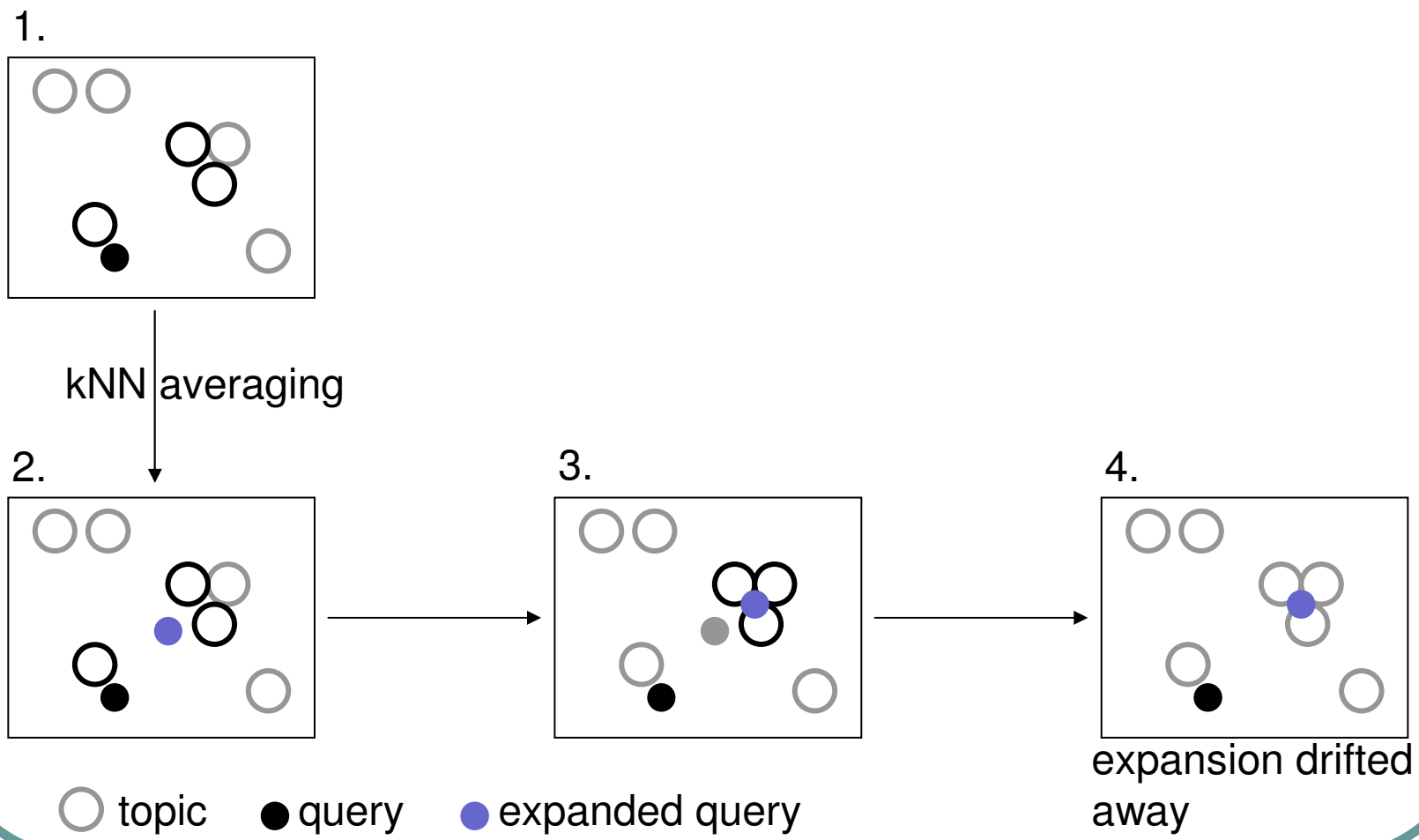
Description: Offers pictures, videos, audios, screen saver, wallpapers, and statistics.

Link: http://www.23jordan.com/

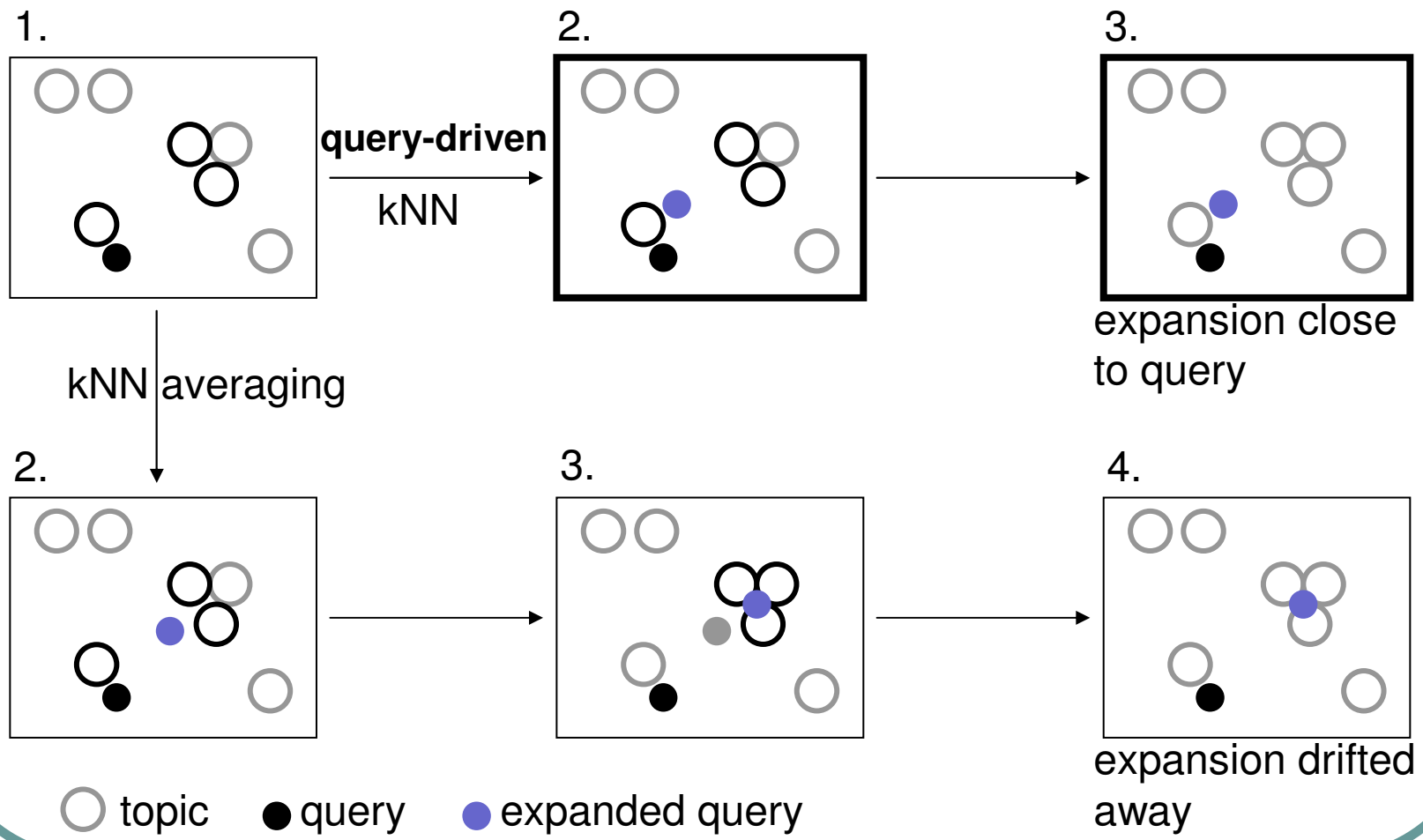
Title: 23Jordan - A Michael Jordan Tribute

Description: News, statistics, pictures, biography, and a forum.

Query Expansion with kNN

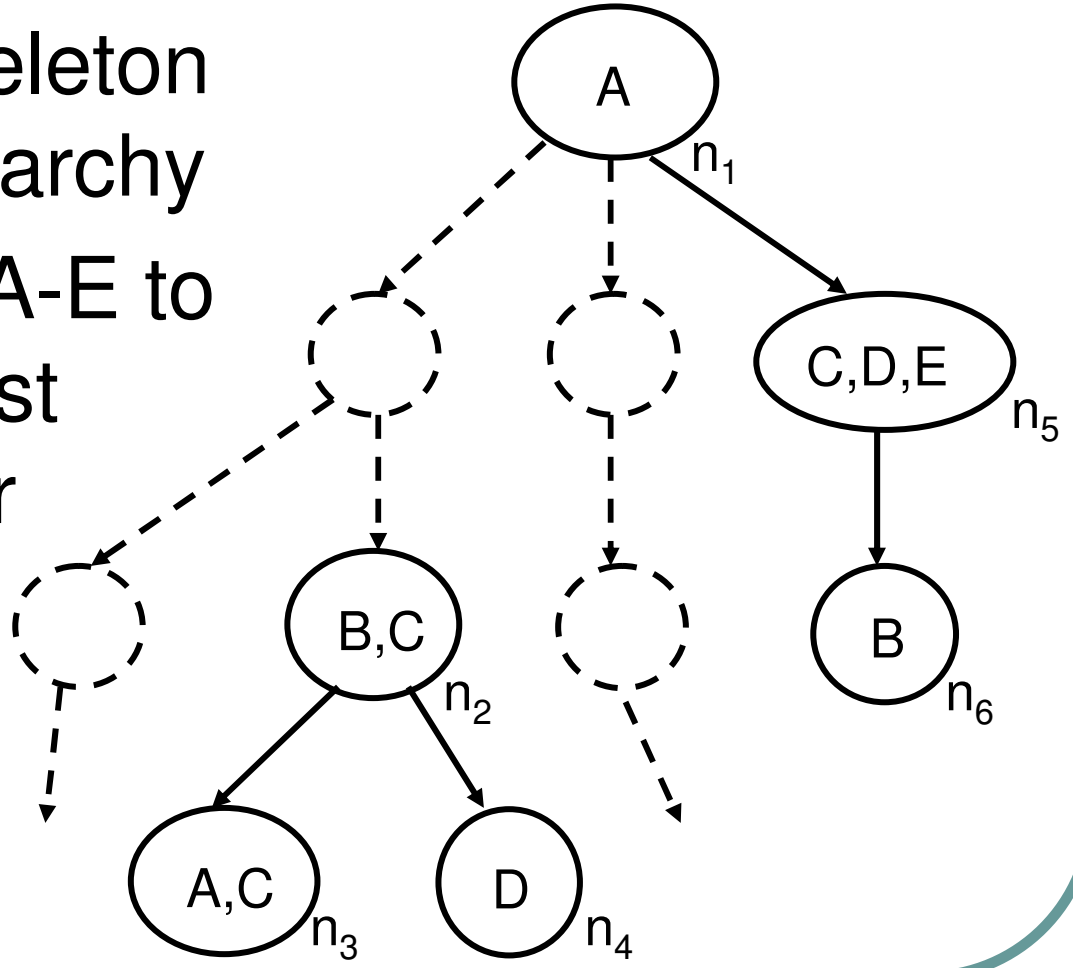


Query Expansion with kNN



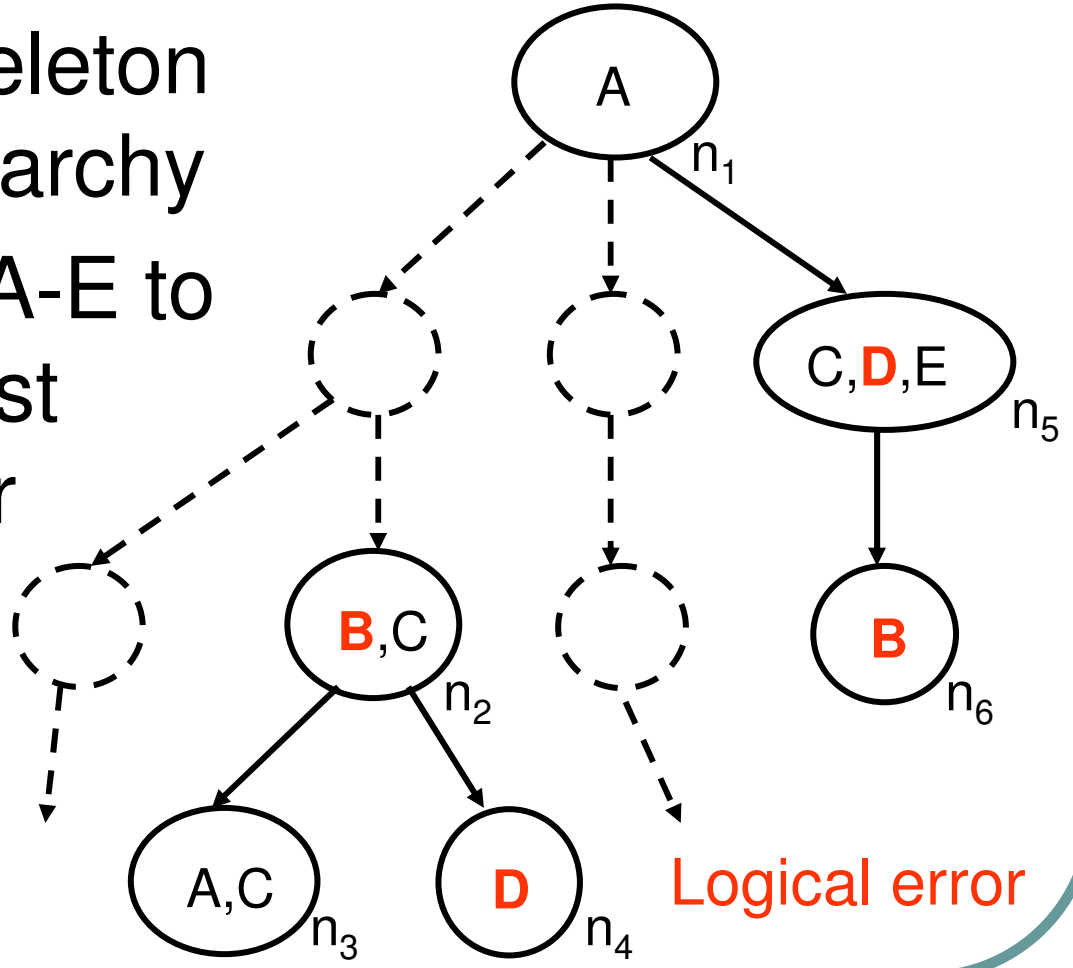
Mapping Queries onto DMOZ

- Take bare skeleton of DMOZ hierarchy
- Map queries A-E to the topics most similar to their expansions



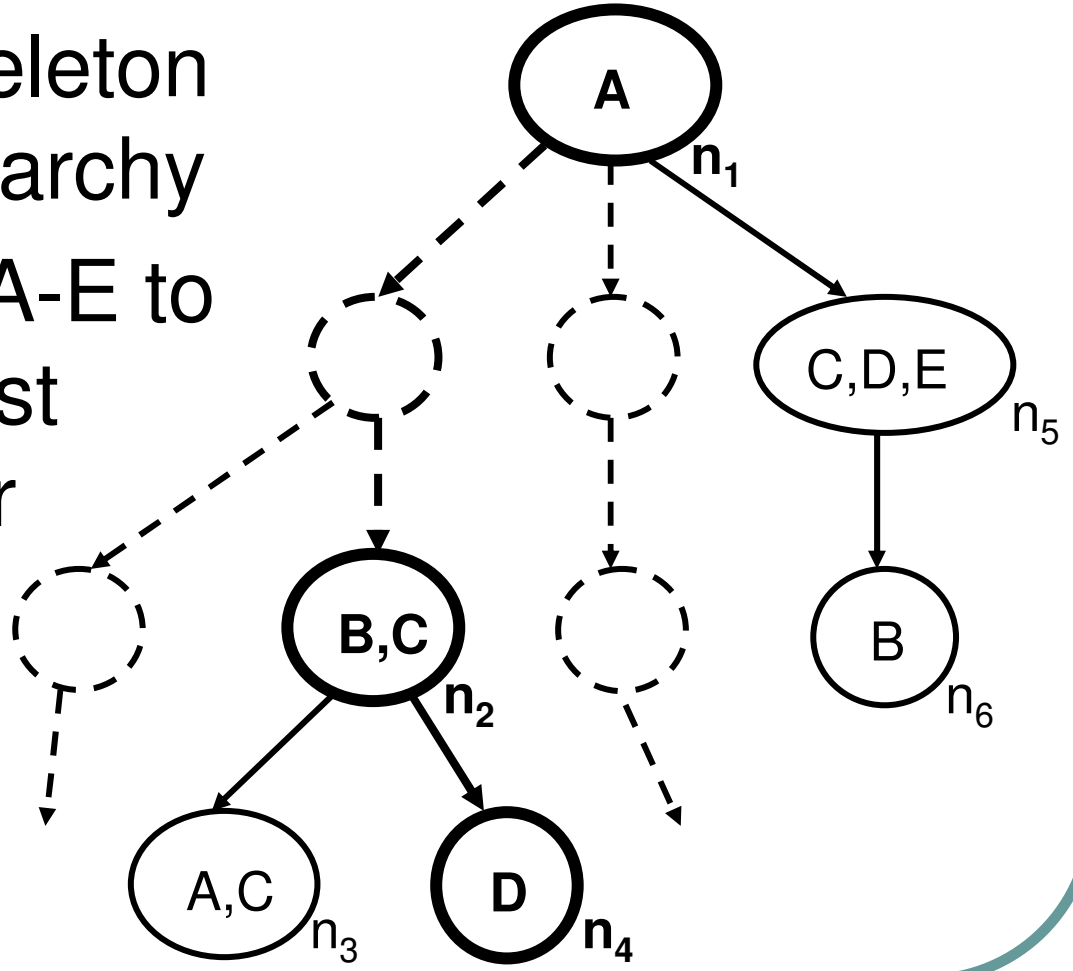
Mapping Queries onto DMOZ

- Take bare skeleton of DMOZ hierarchy
- Map queries A-E to the topics most similar to their expansions

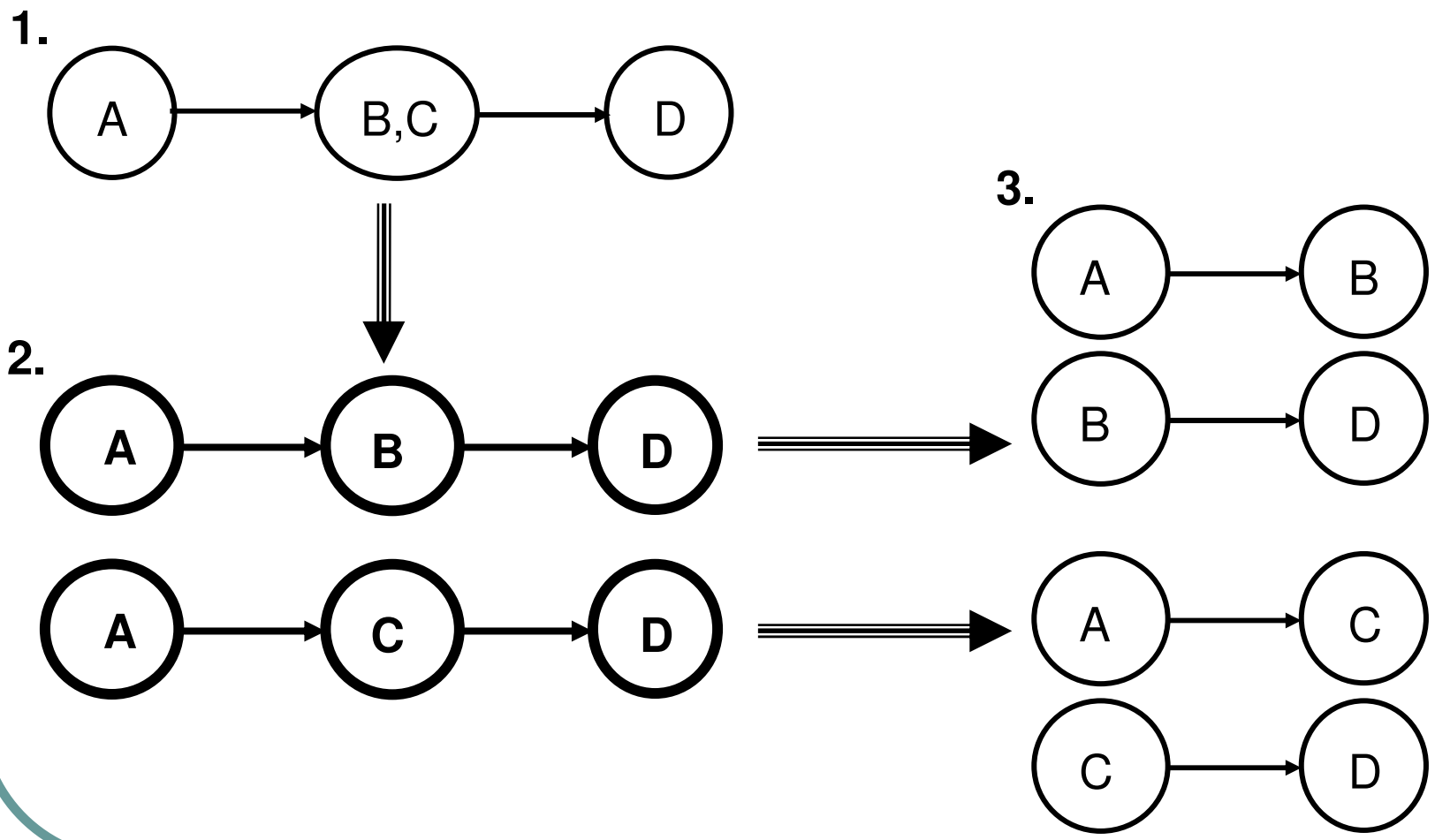


Mapping Queries onto DMOZ

- Take bare skeleton of DMOZ hierarchy
- Map queries A-E to the topics most similar to their expansions

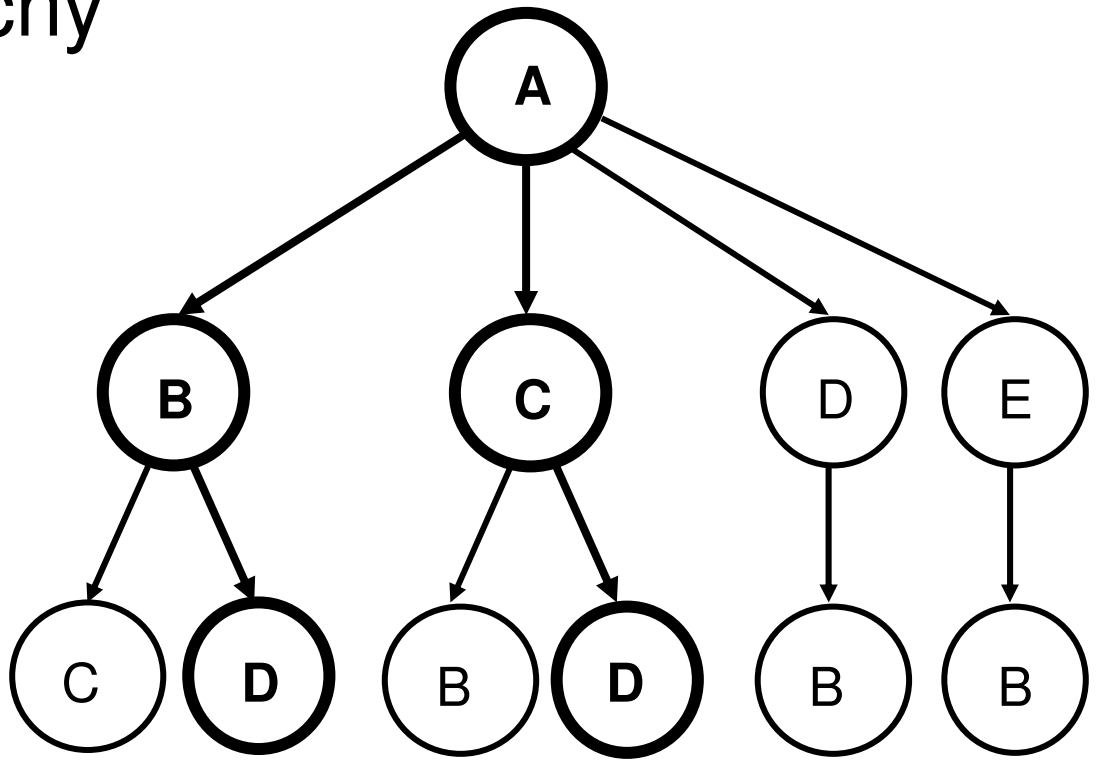


Subsumption Pair Retrieval



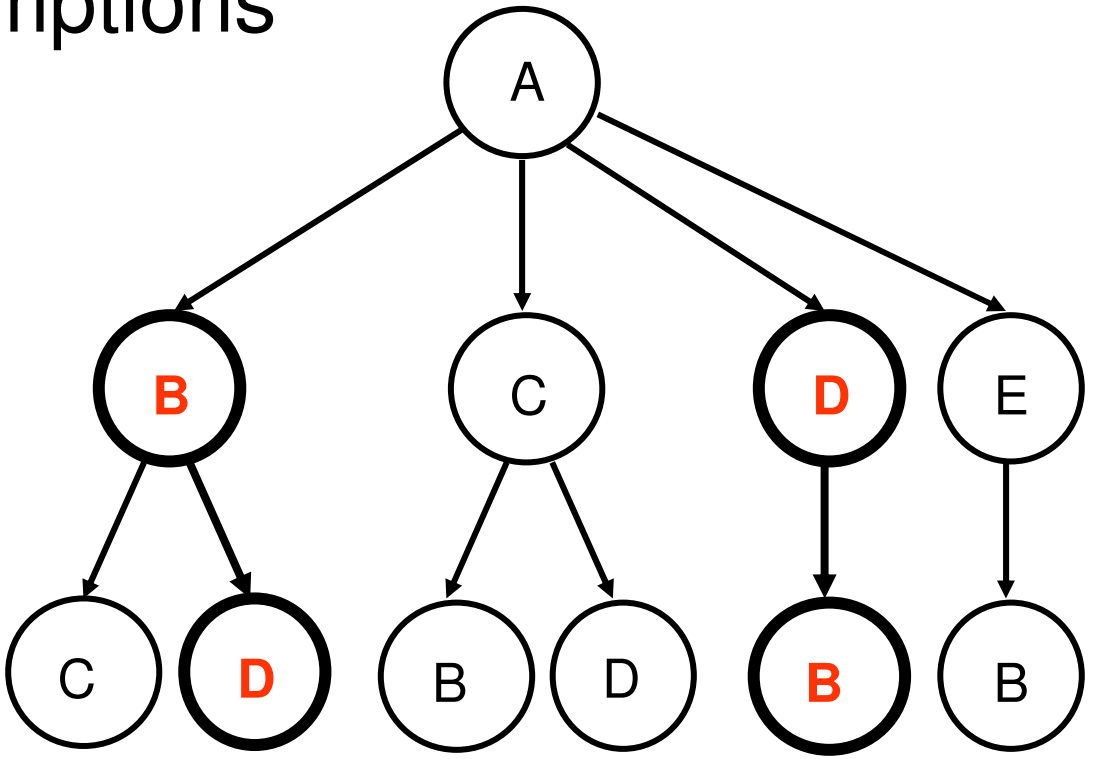
Subsumption Hierarchy Creation

- Combine subsumption pairs and create hierarchy



Subsumption Hierarchy Creation

- Resolve conflicts with context-depending subsumptions



Overview

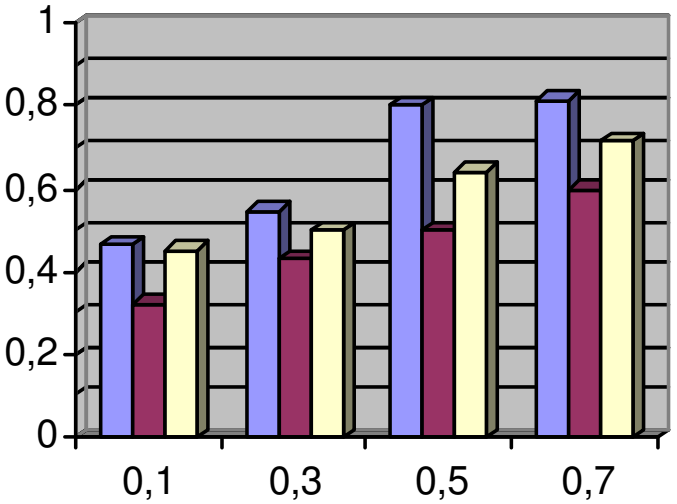
(1) Introduction
(2) Query-Driven K-Nearest Neighbors
(3) Evaluation Results
(4) Application

Query Sets

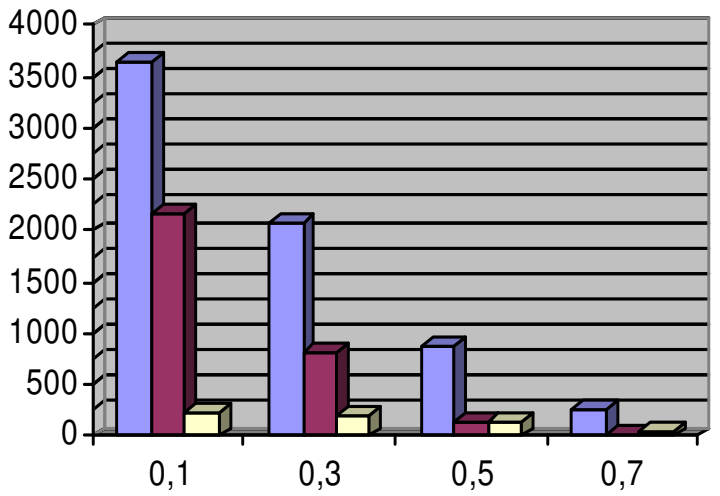
- Hand-crafted query set
  - results not shown here
- 3 sets of 1000 random queries from AOL query log
  - sports-related (“Sports”)
  - health-related (“Health”)
  - domain-unrestricted (“All”)

Evaluation of Q-kNN

- High threshold => high precision
- Low threshold => many subsumption pairs



Precision for different thresholds



Number of subsumption pairs

Different Approaches

- kNN averaging (kNNA)

  query expansion weights all topics equally
- Q-kNN

  weights topics by their similarity to the original query

Different Approaches

- kNN averaging (kNNA)

  query expansion weights all topics equally
- Q-kNN

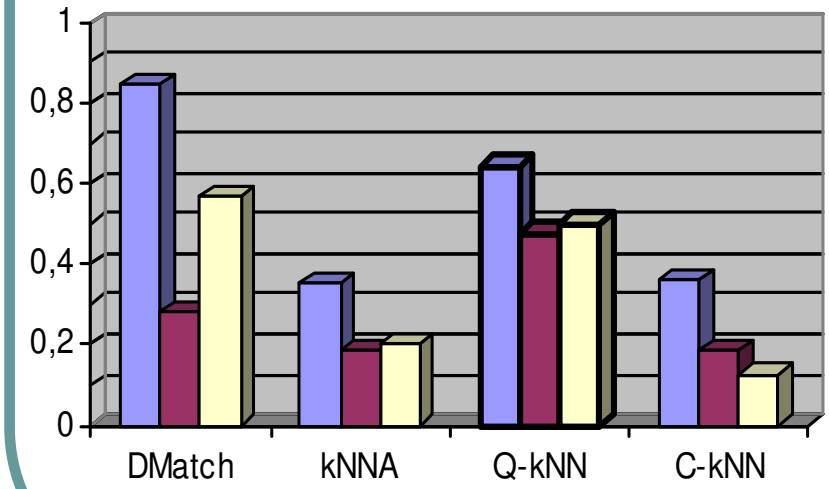
  weights topics by their similarity to the original query
- C-kNN

  weights topics higher which are closer to the query's centroid (expansion)
- Keyword matching (DMatch)

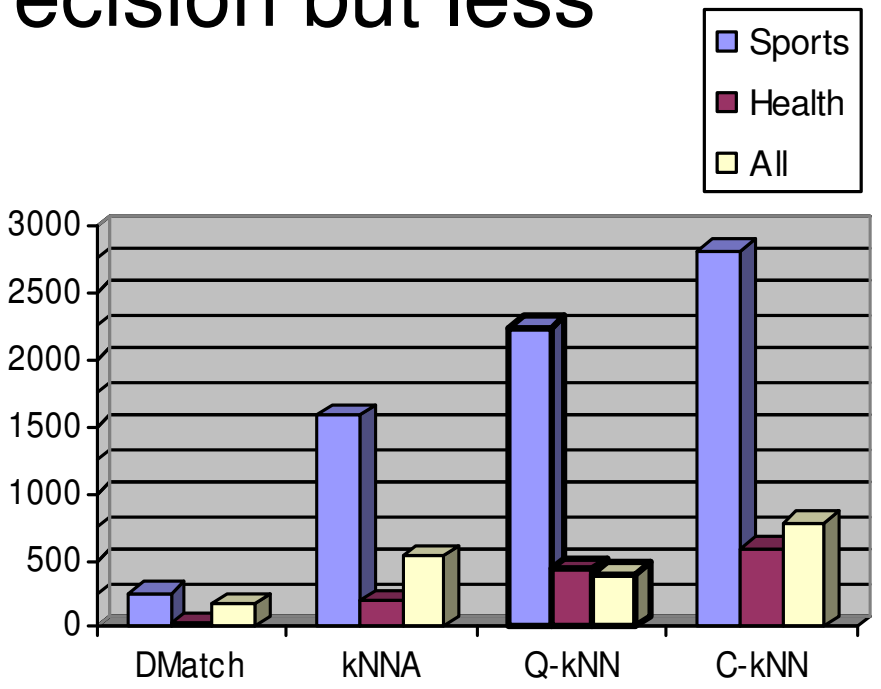
  no query expansion

Comparing Different Approaches

- Q-kNN: high precision for kNN
- DMatch: highest precision but less subsumption pairs

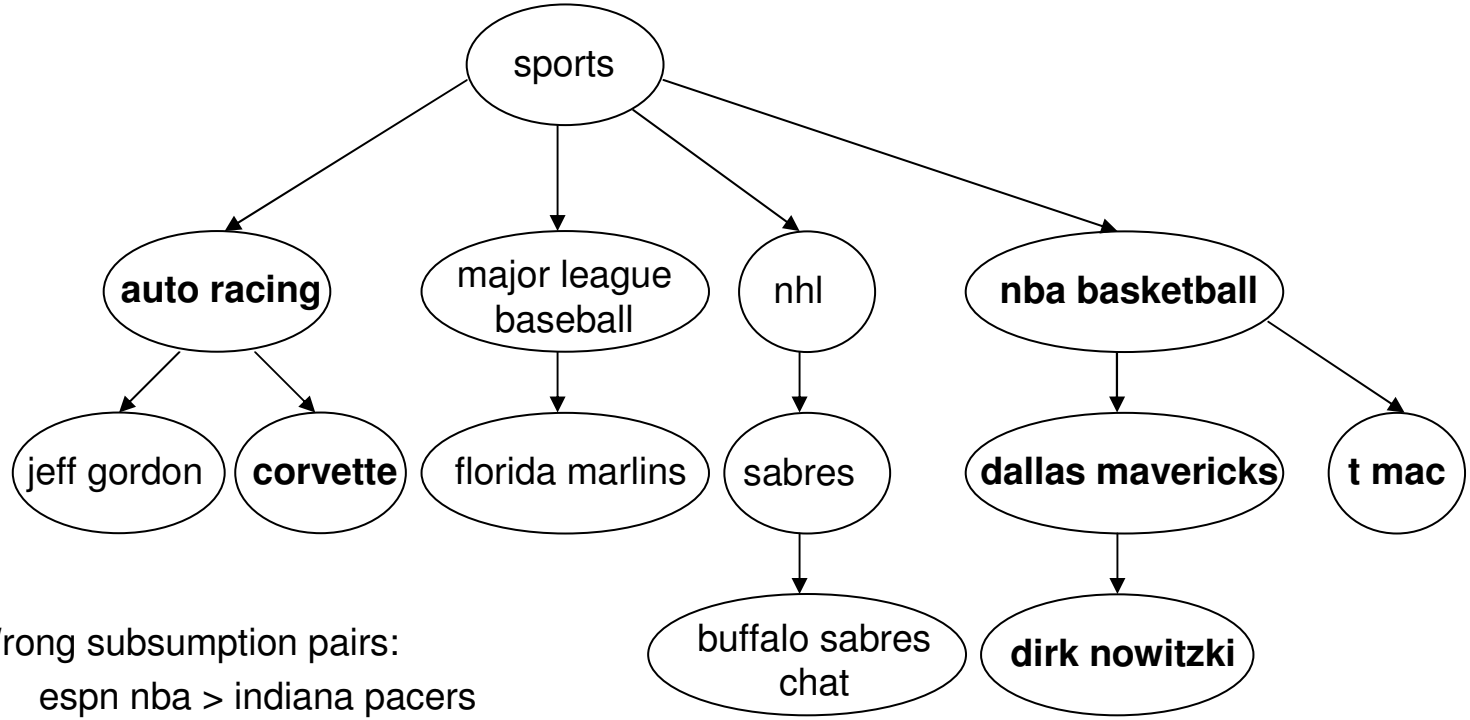


Precision for different thresholds



Number of subsumption pairs

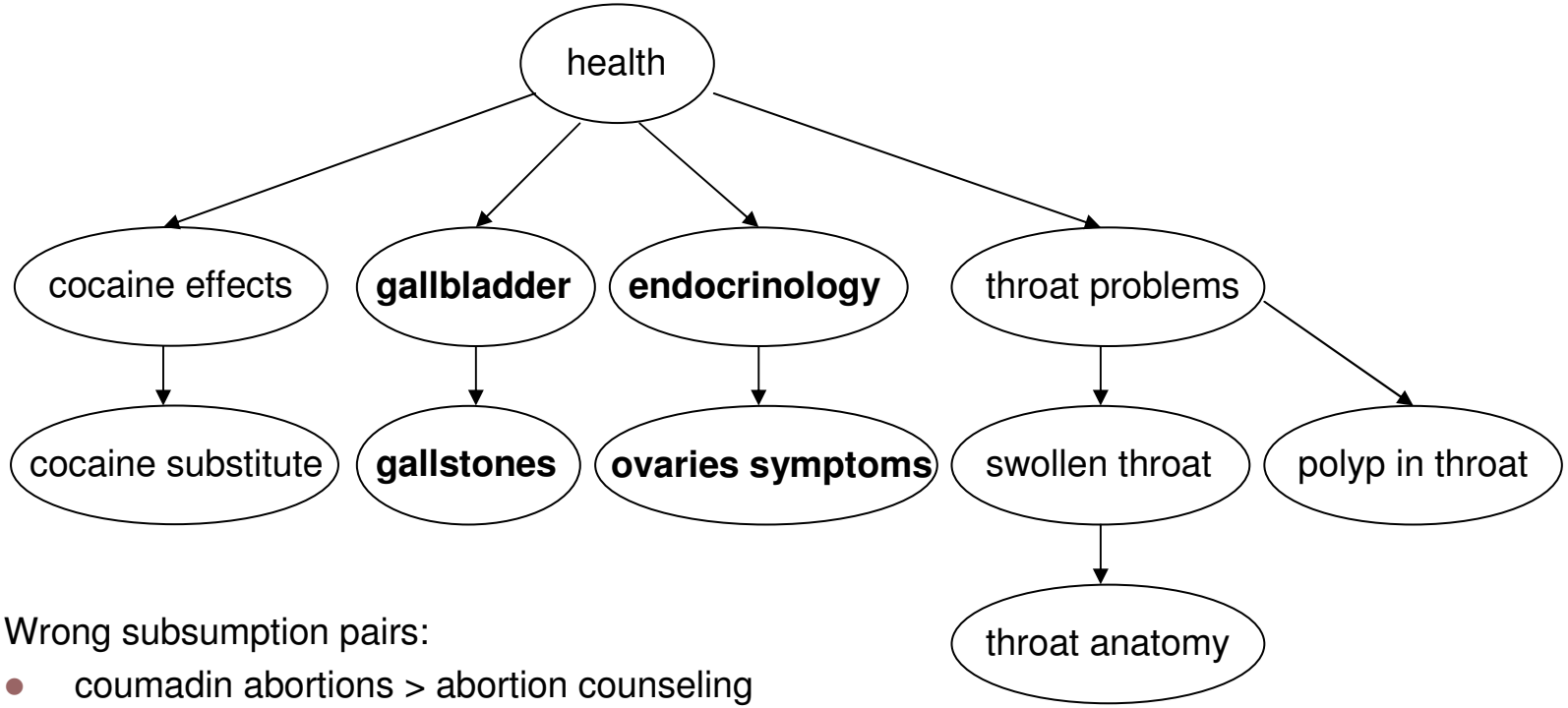
Example Subsumptions (Sports)



Wrong subsumption pairs:

- espn nba > indiana pacers
- ice hockey > msn
- jacksonville jaguars history > jacksonville jaguars
- new england patriots > new england dragway
- photos of pittsburgh steelers > pittsburgh steelers stadium

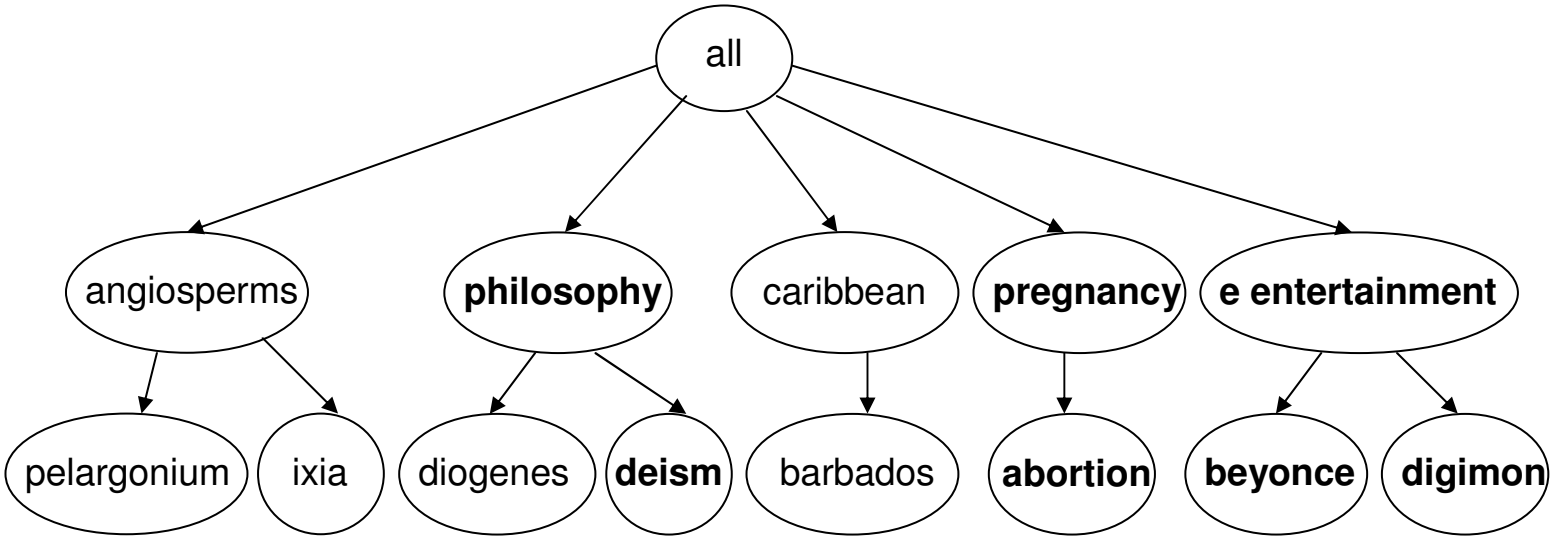
Example Subsumptions (Health)



Wrong subsumption pairs:

- coumadin abortions > abortion counseling
- nursing journals > schoolnurse
- qigong > chinese medicine
- throat tightness > throat problems
- vitamin c foods > calcium magnesium

Example Subsumptions (All)



Wrong subsumption pairs:

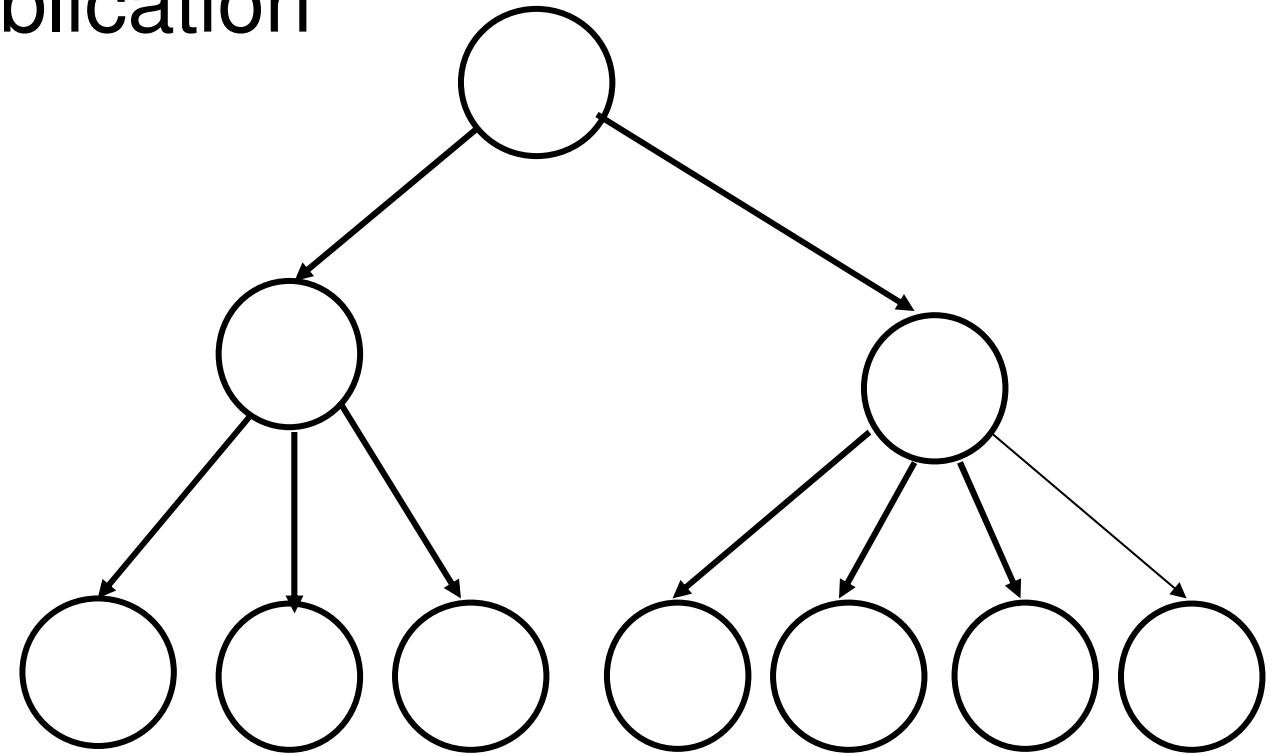
- 1990 > monaco
- angiosperms > africanus
- e trade > dvd-r
- paddle > u.s.s. richmond
- passions > sauna

Overview

(1) Introduction
(2) Query-Driven K-Nearest Neighbors
(3) Evaluation Results
(4) Application

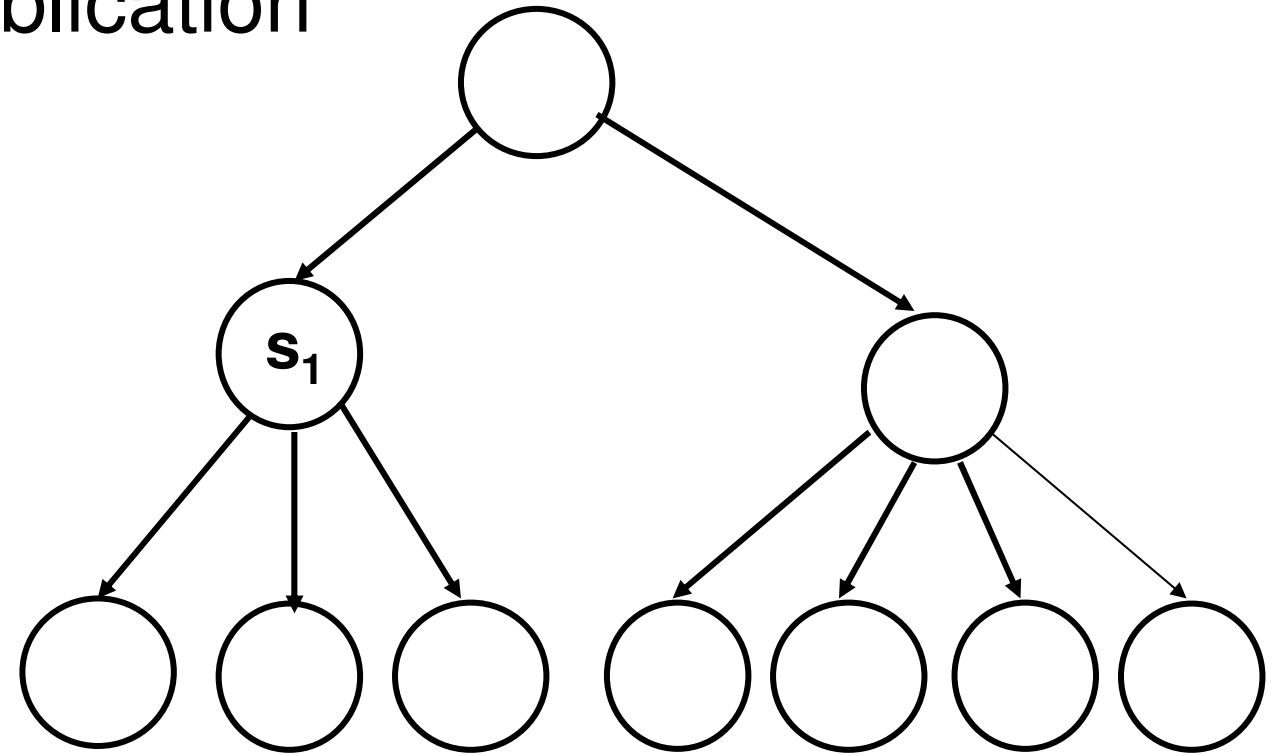
Publish/Subscribe Systems

Fast computation for each subscription and publication



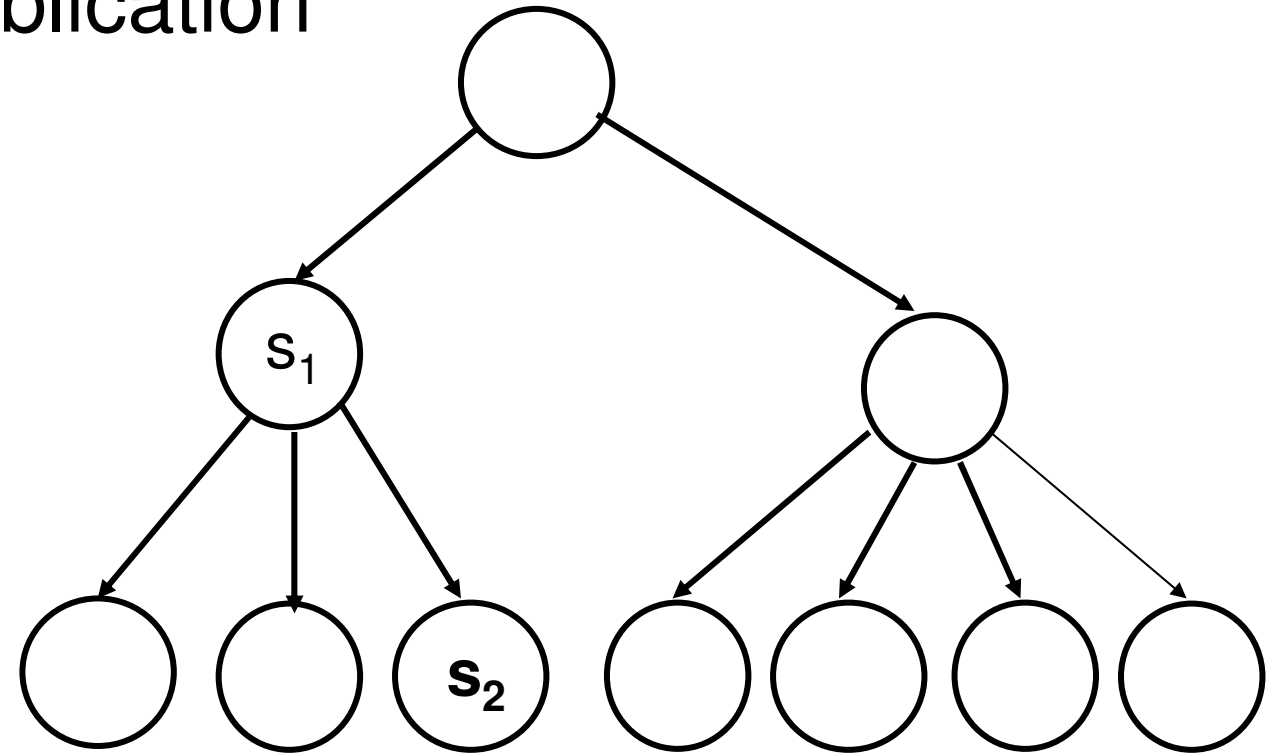
Publish/Subscribe Systems

Fast computation for each subscription and publication



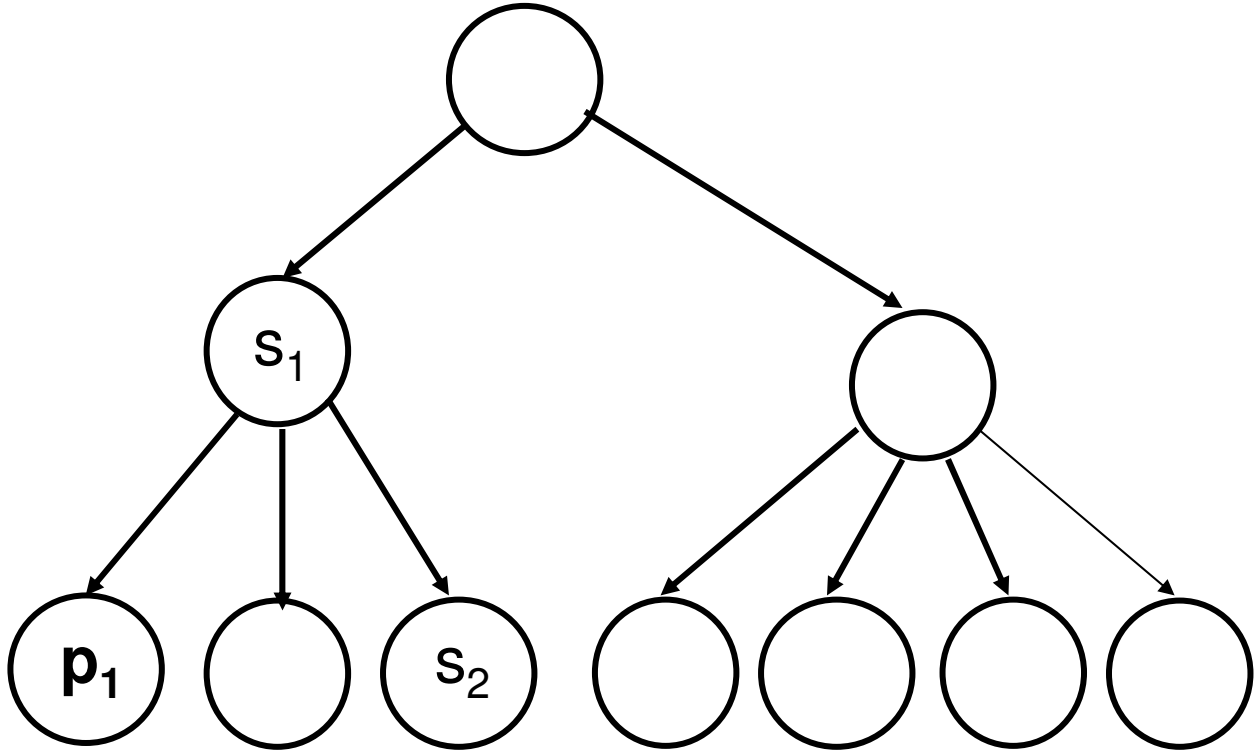
Publish/Subscribe Systems

Fast computation for each subscription and publication



Publish/Subscribe Systems

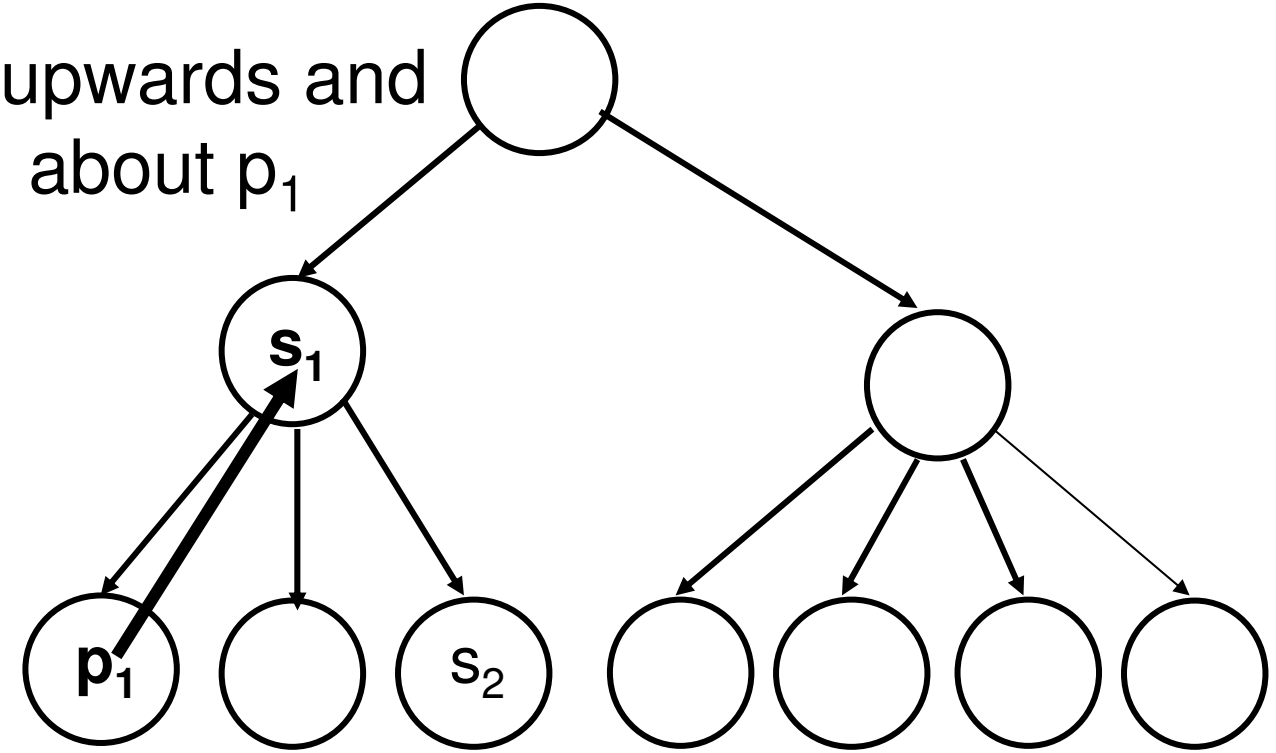
Simple notification about new publications



Publish/Subscribe Systems

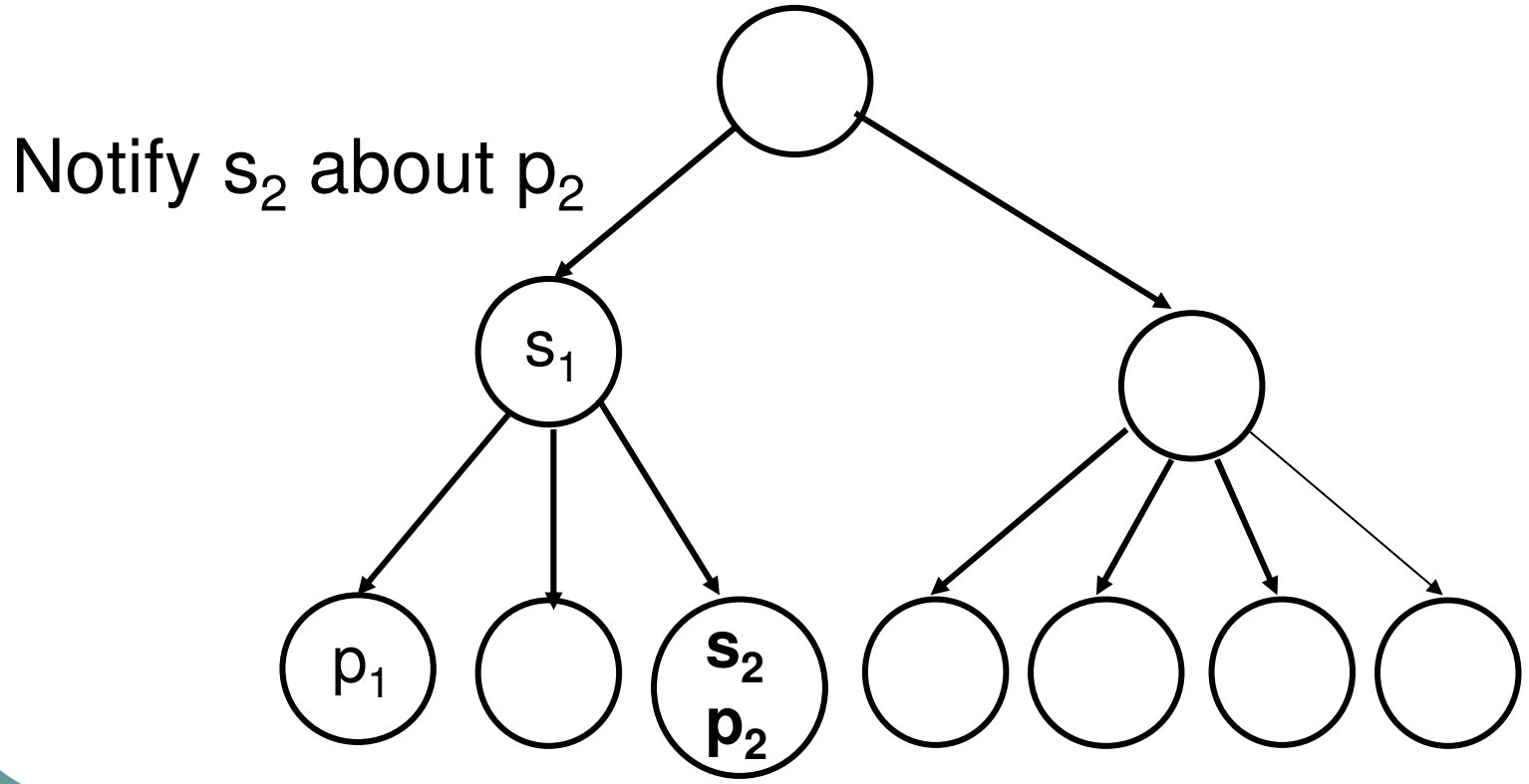
Simple notification about new publications

Search upwards and notify s_1 about p_1



Publish/Subscribe Systems

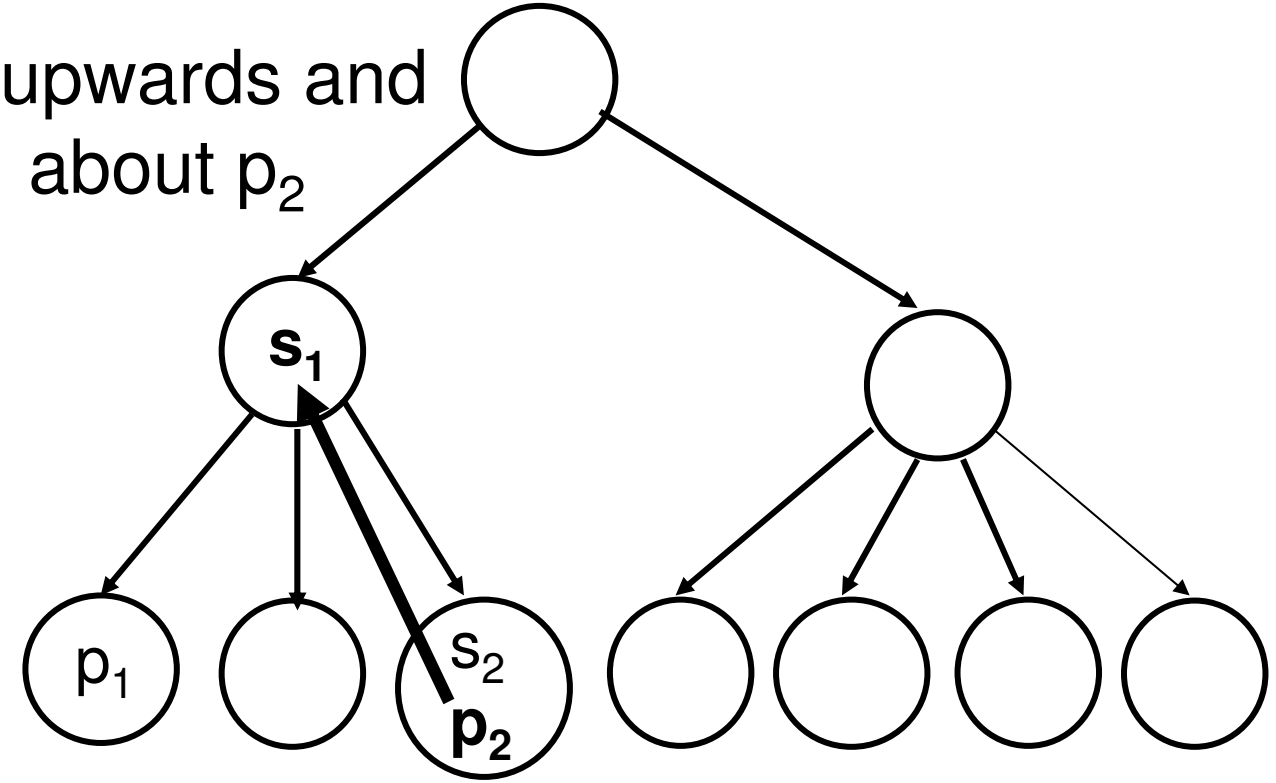
Simple notification about new publications



Publish/Subscribe Systems

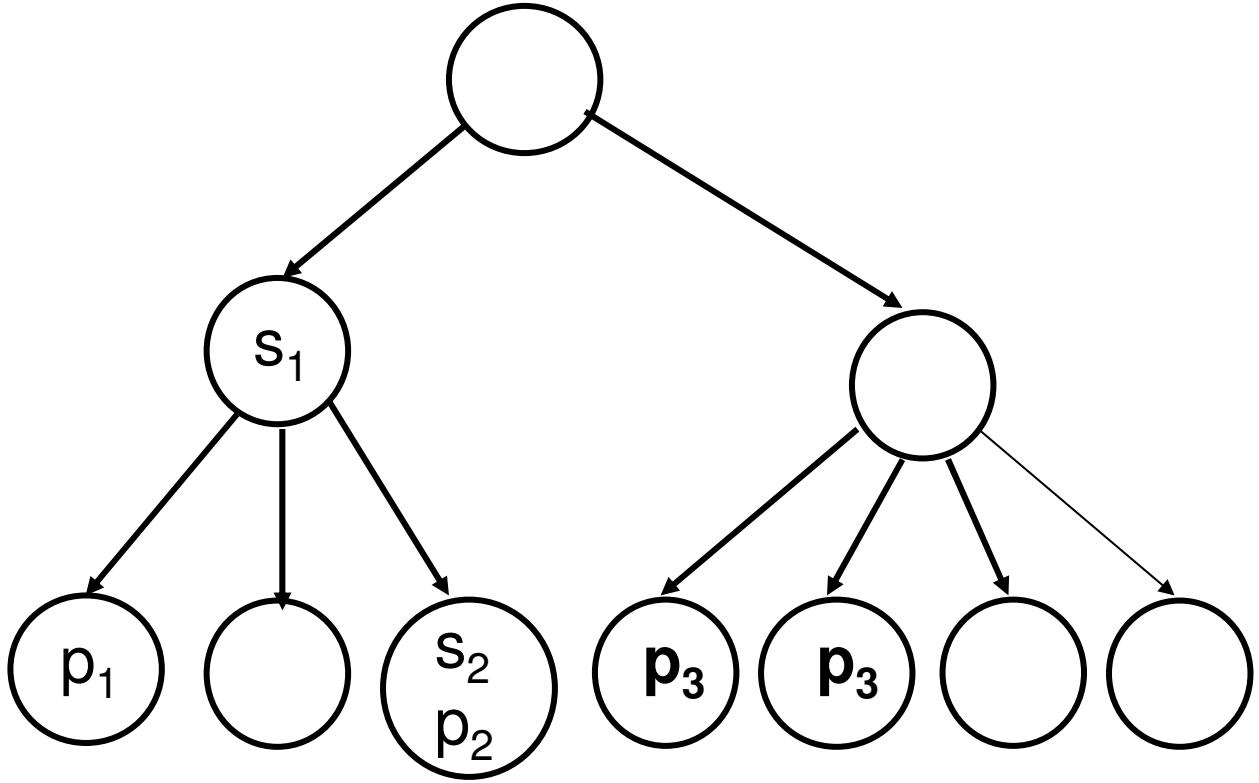
Simple notification about new publications

Search upwards and notify s_1 about p_2



Publish/Subscribe Systems

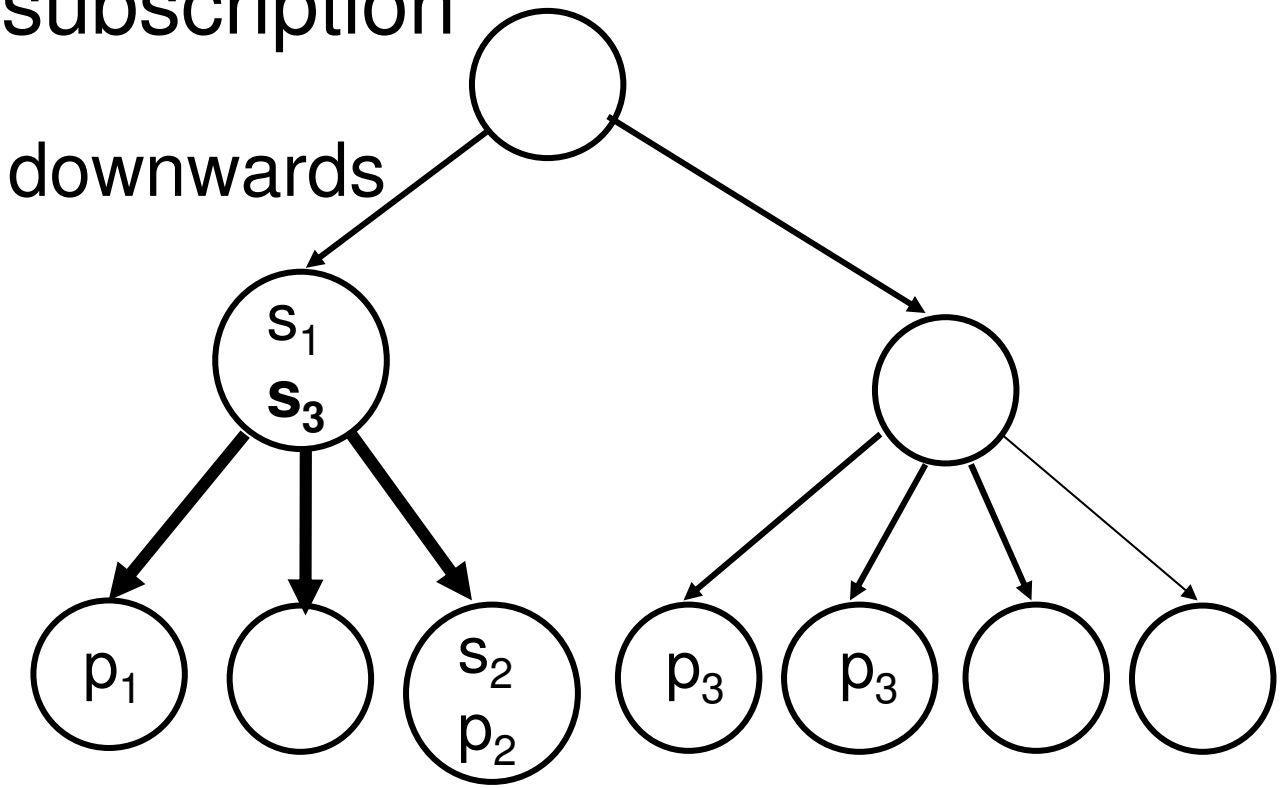
Simple notification about new publications



Publish/Subscribe Systems

Effortless search for publications matching a new subscription

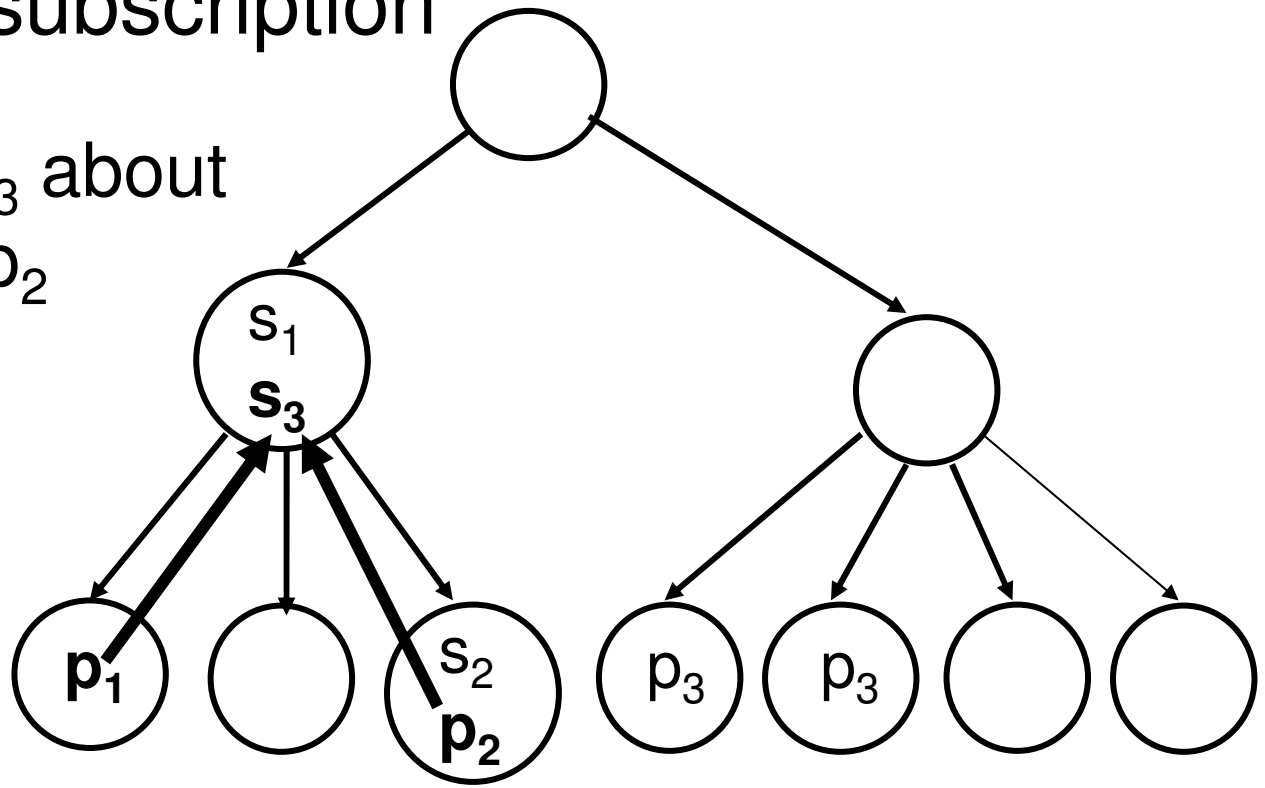
Search downwards



Publish/Subscribe Systems

Effortless search for publications matching a new subscription

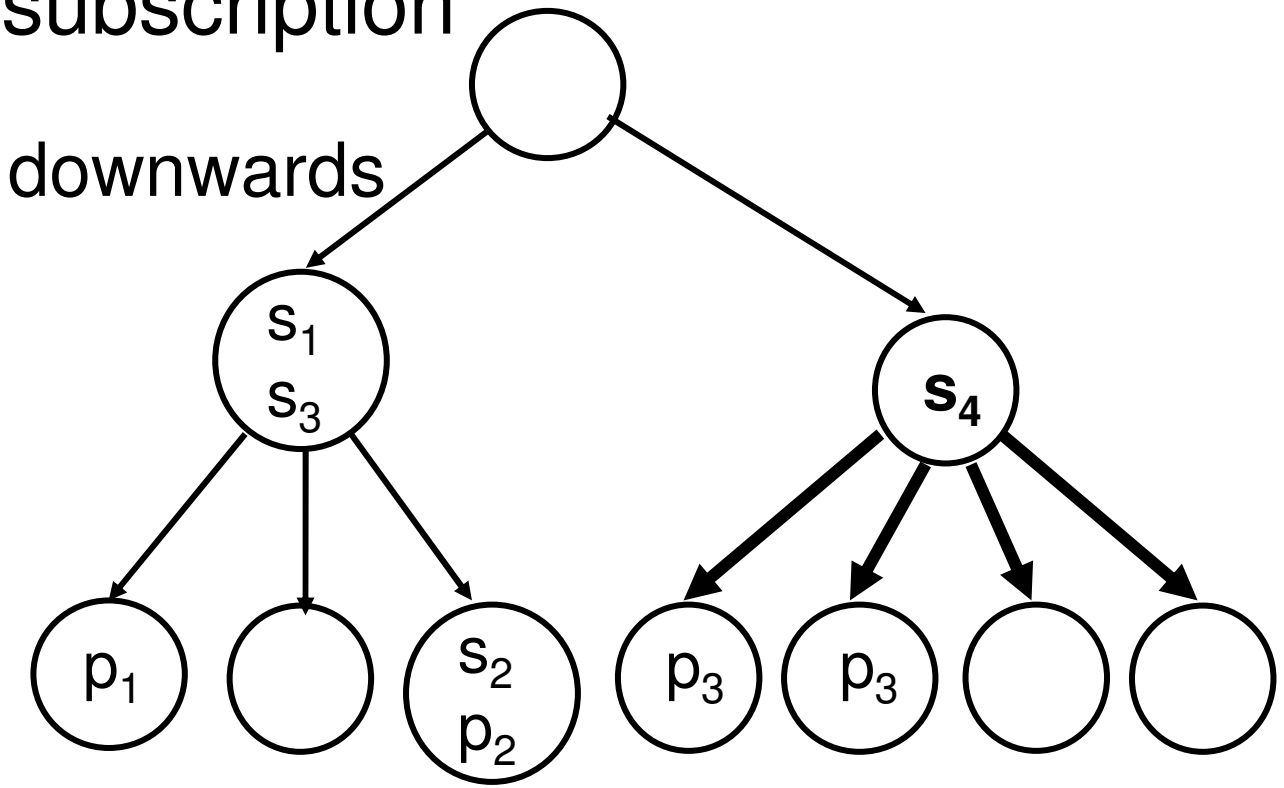
Notify s_3 about p_1 and p_2



Publish/Subscribe Systems

Effortless search for publications matching a new subscription

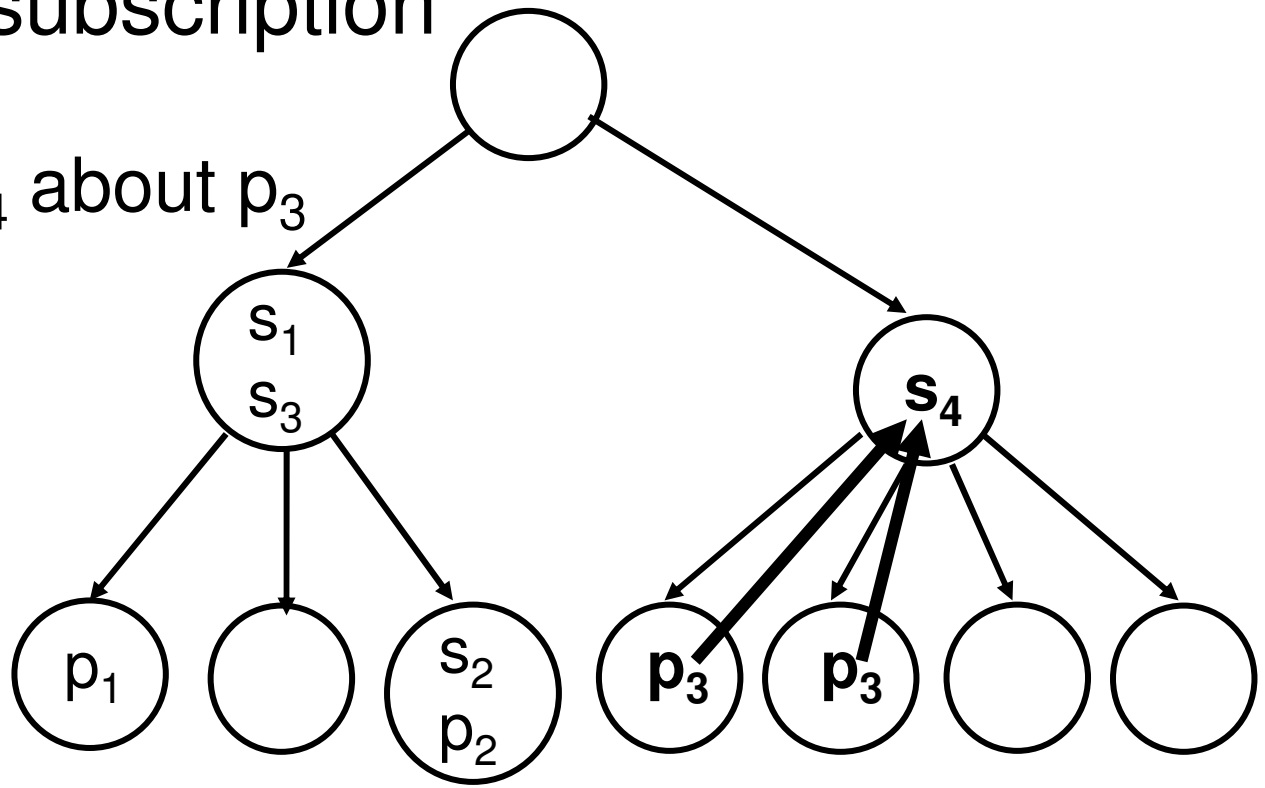
Search downwards



Publish/Subscribe Systems

Effortless search for publications matching a new subscription

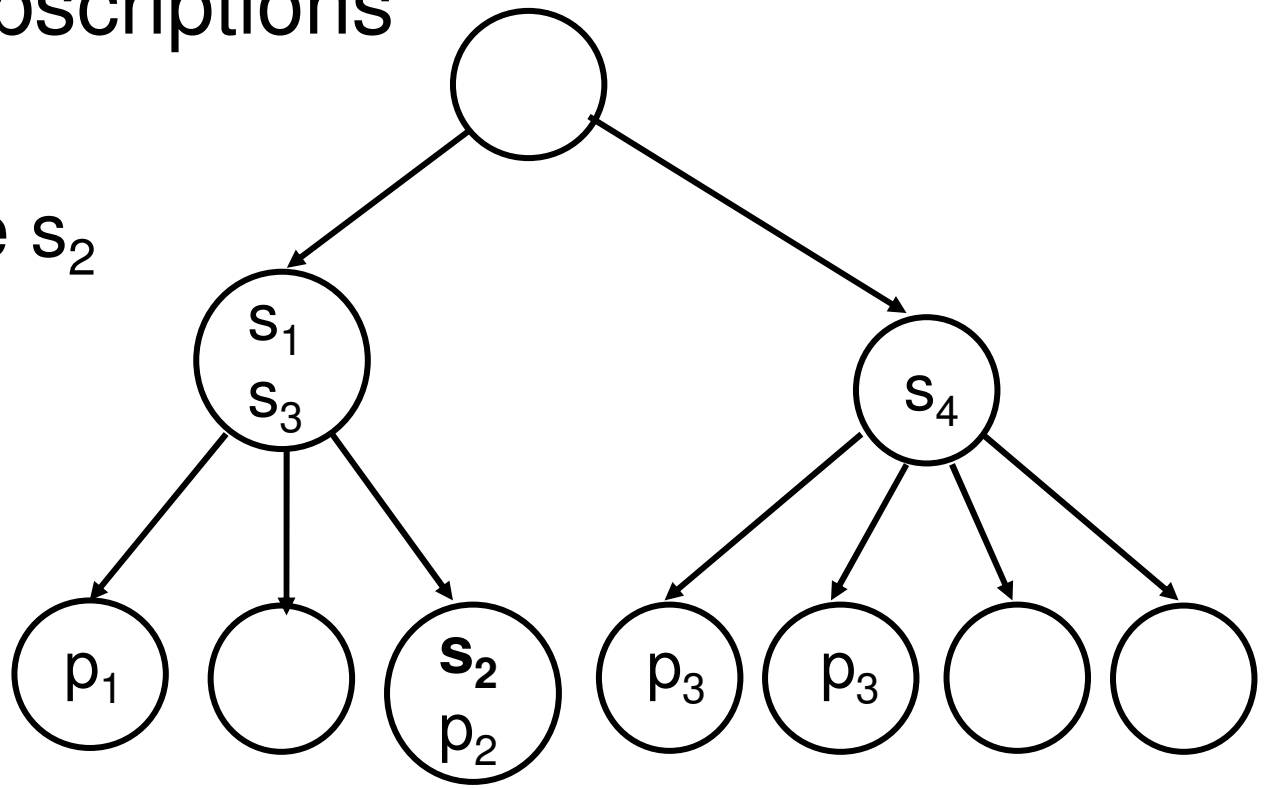
Notify s_4 about p_3



Publish/Subscribe Systems

Straightforward removal of publications and subscriptions

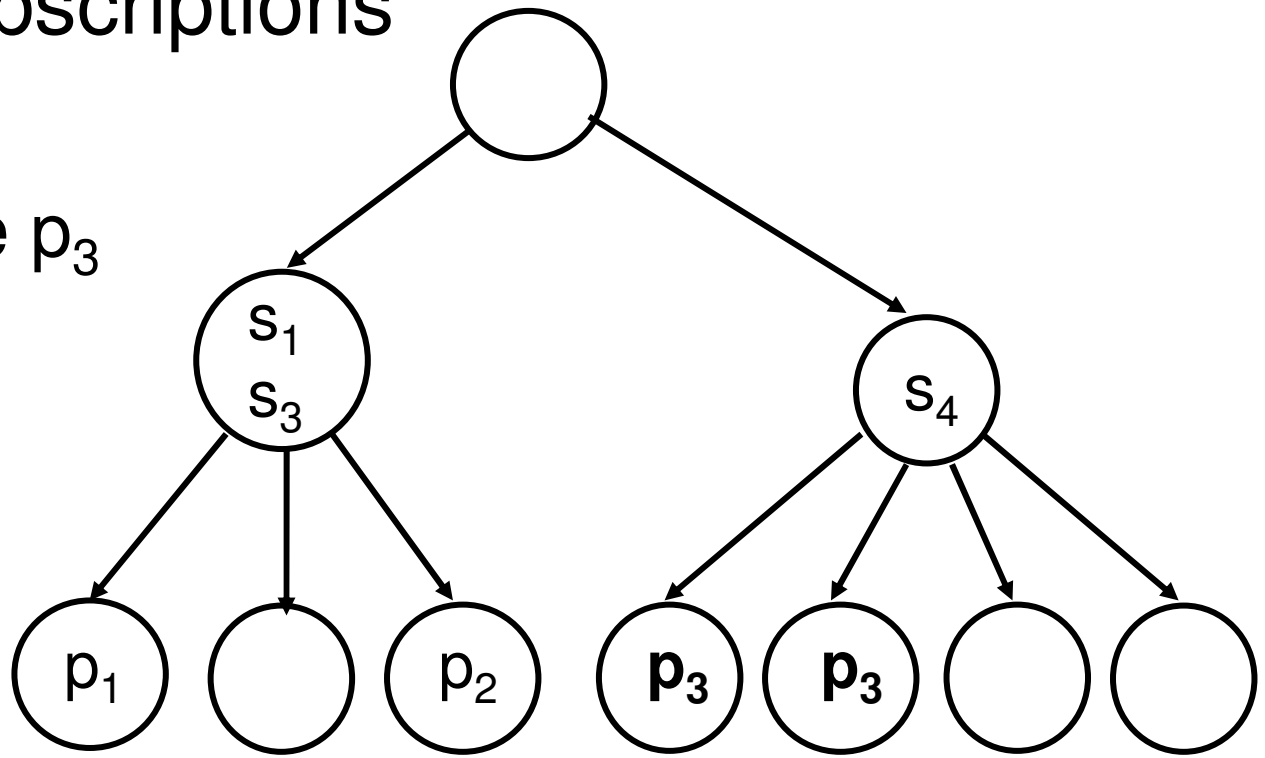
Remove s_2



Publish/Subscribe Systems

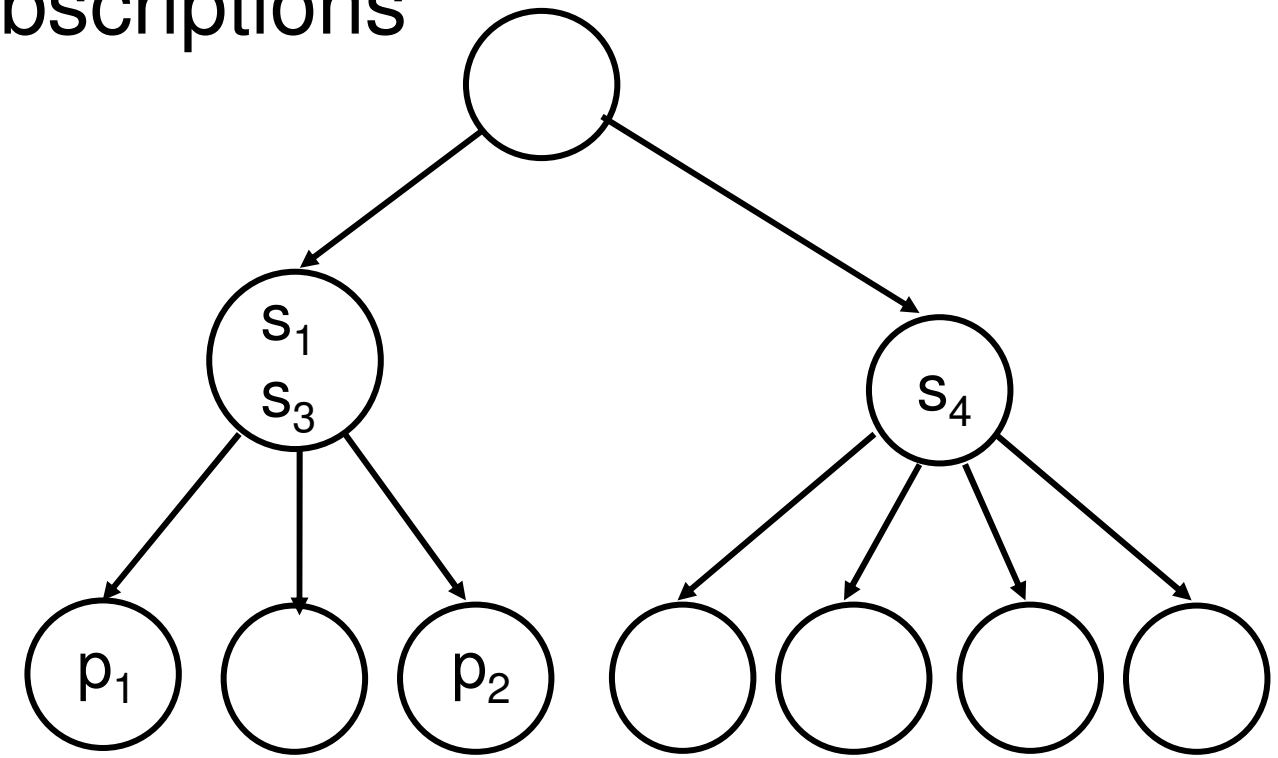
Straightforward removal of publications and subscriptions

Remove p_3



Publish/Subscribe Systems

Straightforward removal of publications and subscriptions



The End

Thank you for your attention.