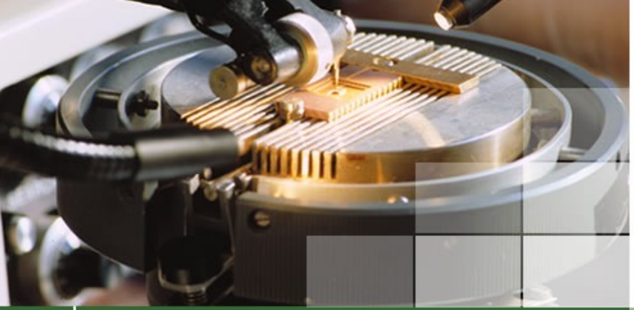


POLITECNICO DI MILANO

Dipartimento di
Elettronica e Informazione

Ranking with uncertain scoring functions

Davide Martinenghi

Joint work with I. Ilyas, M. Soliman, and M. Tagliasacchi

Oxford, May 17, 2011

Summary

- Rank aggregation and rank join
- Uncertain scoring
- Representative orderings
- Sensitivity analysis

Rank aggregation

- Aim: combining several ranked lists of objects in a robust way into a single consensus ranking
  - Objects are equipped with a score
  - An aggregation function computes the overall score
    - Typically monotone (e.g., weighted sum)
- Main interest in the top k elements of the aggregation
  - Need for algorithms that quickly obtain the top results
  - … without having to read each ranking in its entirety
- Data access is sorted (from top scores downwards)
  - Some works also allow random access: given an object, retrieve its score

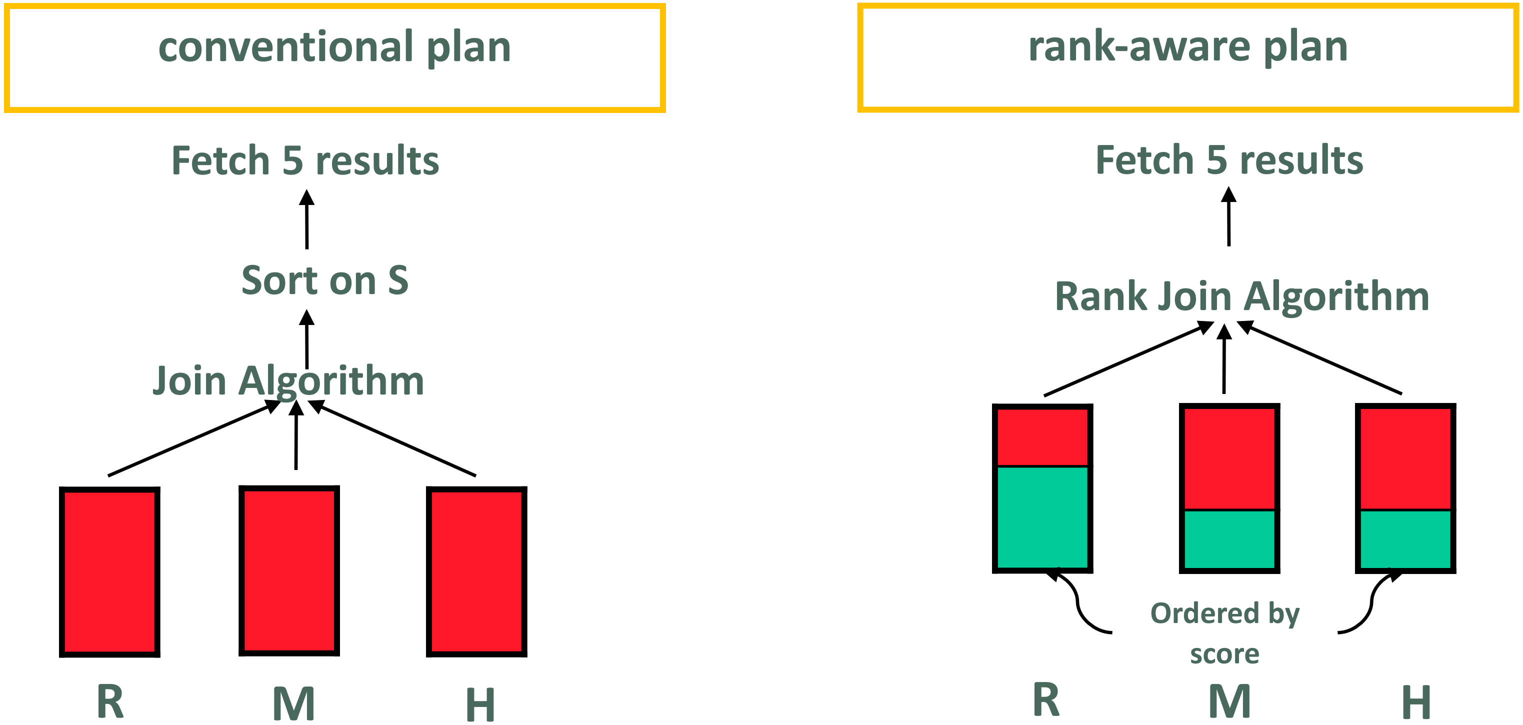
Rank join

- Extends rank aggregation to different data sets
  - A natural join $R_1 \bowtie R_2 \dots \bowtie R_n$
  - A scoring function $\mathcal{S}(\tau) = f\left(\mathcal{S}(\tau_1), \dots, \mathcal{S}(\tau_n)\right)$
  - A positive integer $k < |R_1 \bowtie R_2 \dots \bowtie R_n|$
- Compute
  - k join results with highest scores

[Ilyas et al., VLDB2004]
[Schnaitter and Polyzotis, PODS2008]

Rank-aware plans


```
SELECT r.id, m.id, h.id,
FROM RestaurantsNY r, MovieTheatersNY m, HotelsNY h,,
WHERE r.neighborhood = h.neighborhood = m.neighborhood
RANK BY 0.5*r.price + 0.3*m.rating + 0.2*h.stars
LIMIT 5
```



[Adapted from Polyzotis, 2010]

Inspiring work

- Soliman and Ilyas, “Ranking with uncertain scores”, ICDE 2009
  - Objects have scores defined over intervals
    - E.g., apartment rent [$200-$250]
- Vlachou et al. “Reverse Top-k queries”, ICDE 2010
  - Given a set of (linear) scoring functions, determine the one that gives the highest rank for a target object

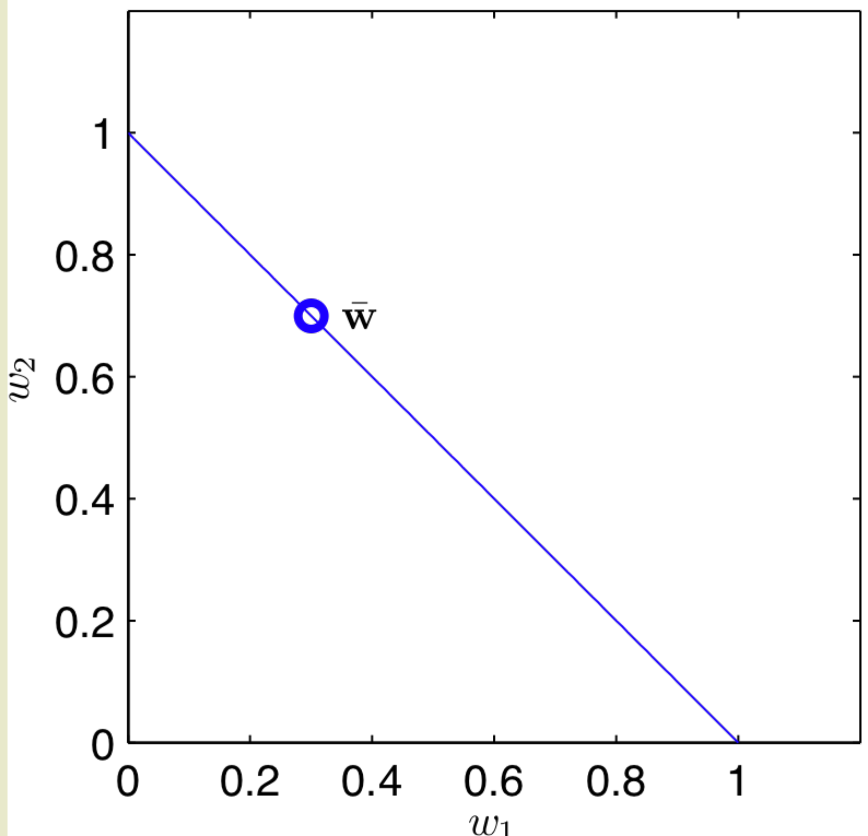
Uncertain scoring

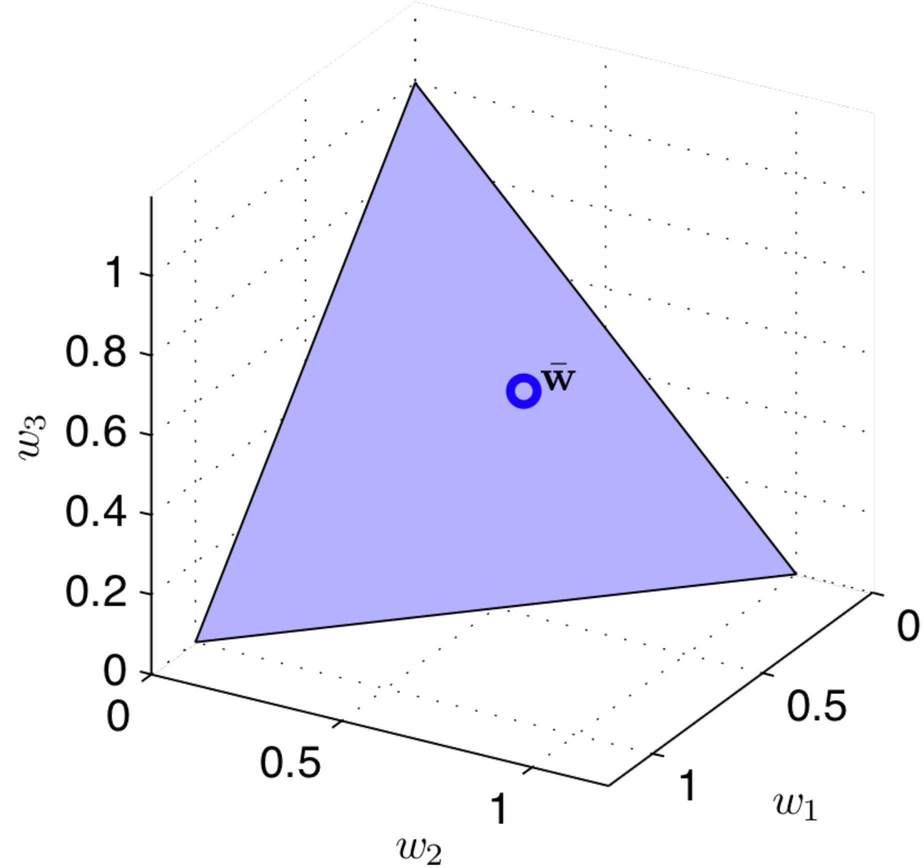
- Users are often unable to precisely specify the scoring function
- Using trial-and-error or machine learning may be tedious and time consuming
- Even when the function is known, it is crucial to analyze the sensitivity of the computed ordering wrt. changes in the function

Uncertain scoring

- Assumptions:
  - Linear scoring function
    - $S = w_1s_1 + w_2s_2 + \ldots + w_ns_n$
  - User-defined weights $w_1, w_2, \ldots, w_n$ are
    - Uncertain, and, w.l.o.g.,
    - normalized to sum up to 1

Representing scoring functions on the simplex





- Each point on the simplex represents a possible scoring function
- We assume that p(**w**) is uniform over the simplex

Uncertain scoring

- Uncertainty induces a probability distribution on a set of possible orderings
- Each ordering occurs with a probability

$$p(\boldsymbol{\lambda}_N) = \int_{\mathbf{w} \in \Delta^{d-1}, \mathcal{O} \overset{\mathbf{w}}{\rightsquigarrow} \boldsymbol{\lambda}_N} p(\mathbf{w}) d\mathbf{w}$$

(weights in the simplex inducing that ordering)

- When N is large, we usually focus on a prefix of length K<N of an ordering

Example

- Top-k query:

```
SELECT R.RestName, R.Street, H.HotelName
FROM RestaurantsInParis R, HotelsInParis H
WHERE distance(R.coordinates, H.coordinates) ≤ 500m
RANK BY w_R· R.Rating + w_H· H.Stars
LIMIT 5
```

- Results and possible orderings:

ID	rating	stars
τ_1	2	6
τ_2	7	5
τ_3	4	7
τ_4	5	2

Join Results

Rank By $w_R.rating+w_H.stars$

$w_R+w_H=1$

λ^1	λ^2	λ^3	λ^4	λ^5
τ_3	τ_3	τ_2	τ_2	τ_2
τ_1	τ_2	τ_3	τ_3	τ_4
τ_2	τ_1	τ_1	τ_4	τ_3
τ_4	τ_4	τ_4	τ_1	τ_1

w_R: 0 0.167 0.4 0.571 0.833 1.0

Objectives of our study (1/2)

- Finding a representative ordering:
  - Most Probable Ordering:

$$\boldsymbol{\lambda}^{*}_{MPO} = arg. \max_{\boldsymbol{\lambda} \in \Lambda_K} p(\boldsymbol{\lambda})$$

  - Optimal Rank Aggregation:
    - Ordering with the minimum average distance to all other orderings
  - Common distances between orderings:
    - Kendall tau: number of pairwise disagreements in the relative order of items
    - Spearman's footrule: sum of distances between the ranks of the same item in the two orderings

Example of MPO and ORA

- For K=2, the MPO is $<\tau_2, \tau_3>$
- ORA is λ^3 both for Kendall tau and footrule

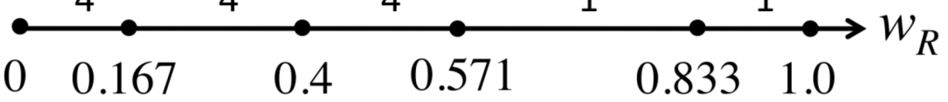
ID	rating	stars
τ_1	2	6
τ_2	7	5
τ_3	4	7
τ_4	5	2

Join Results

Rank By $w_R.rating+w_H.stars$

$w_R+w_H=1$

λ^1	λ^2	λ^3	λ^4	λ^5
τ_3	τ_3	τ_2	τ_2	τ_2
τ_1	τ_2	τ_3	τ_3	τ_4
τ_2	τ_1	τ_1	τ_4	τ_3
τ_4	τ_4	τ_4	τ_1	τ_1



Objectives of our study (2/2)

- Quantifying sensitivity
  - Stability of a chosen ordering wrt. perturbations in the weights
    - largest volume in the weights space, around an input weight vector **w**, in which changing the weights leaves the computed ordering unaltered
  - Likelihood of a chosen ordering
    - probability of obtaining an ordering identical to a given one up to depth K

Example of Stability

- For **w**=(0.2,0.8) we have λ^2
- For K=2, the vector (.167,.833) is the furthest that still induces λ^2
- The measure of stability is the distance
  $||(0.2,0.8) - (.167,.833)||=0.047$

ID	rating	stars
τ_1	2	6
τ_2	7	5
τ_3	4	7
τ_4	5	2

Join Results

Rank By $w_R.rating+w_H.stars$

$w_R+w_H=1$

λ^1	λ^2	λ^3	λ^4	λ^5
τ_3	τ_3	τ_2	τ_2	τ_2
τ_1	τ_2	τ_3	τ_3	τ_4
τ_2	τ_1	τ_1	τ_4	τ_3
τ_4	τ_4	τ_4	τ_1	τ_1

w_R: 0, 0.167, 0.4, 0.571, 0.833, 1.0

Example of Likelihood

- For **w**=(0.5,0.5) we have λ^3
- For K=2, likelihood is

$$p(\lambda^3) + p(\lambda^4)$$

(λ^3 and λ^4 identical up to depth 2)

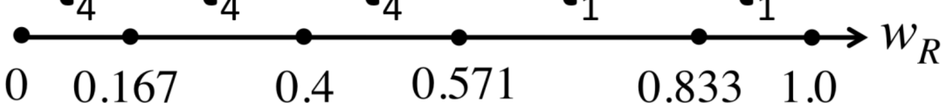
ID	rating	stars
τ_1	2	6
τ_2	7	5
τ_3	4	7
τ_4	5	2

Join Results

Rank By $w_R.rating+w_H.stars$

$w_R+w_H=1$

λ^1	λ^2	λ^3	λ^4	λ^5
τ_3	τ_3	τ_2	τ_2	τ_2
τ_1	τ_2	τ_3	τ_3	τ_4
τ_2	τ_1	τ_1	τ_4	τ_3
τ_4	τ_4	τ_4	τ_1	τ_1



Computing representative orderings

- A naïve approach:
  1. Enumerate possible weight vectors
  2. Find the distinct orderings induced by these vectors
  3. Pick the required representative ordering

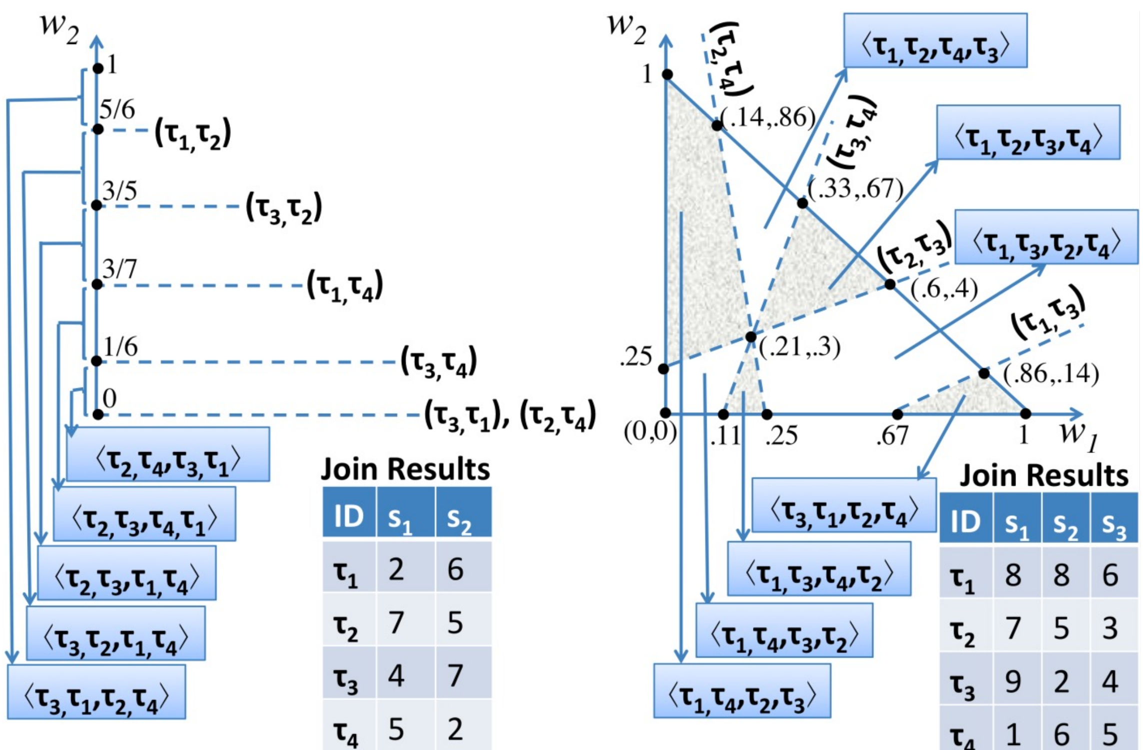
- This is:
  - Highly inefficient
  - Inaccurate, since it requires discretizing the weights space

Efficient approaches

- MPO requires processing prefixes
- ORA requires processing full orderings
- A holistic approach: succinct representation of full orderings as disjoint partitions of the space of weights
  - Appropriate for ORA
- An incremental approach: tree-based representation that is incrementally constructed by extending prefixes of orderings
  - Appropriate for MPO

Holistic approach

- For each pair of join results Ti and Tj
  - Divide the space of weights into two partitions based on their aggregate score
    - In one F(Ti)>F(Tj), in the other F(Ti)<F(Tj)
  - The space is thus partitioned into $O(N^{2^{(d-1)}})$ disjoint convex polyhedra, each corresponding to an ordering



Join Results

ID	s_1	s_2
τ_1	2	6
τ_2	7	5
τ_3	4	7
τ_4	5	2

(a)

Join Results

ID	s_1	s_2	s_3
τ_1	8	8	6
τ_2	7	5	3
τ_3	9	2	4
τ_4	1	6	5

(b)

Finding ORA using the holistic approach

- ORA under Kendall tau for d=2
  - Simply given by sorting join results using the sum of the score components as the sort comparator
  - Uses weak stochastic transitivity:

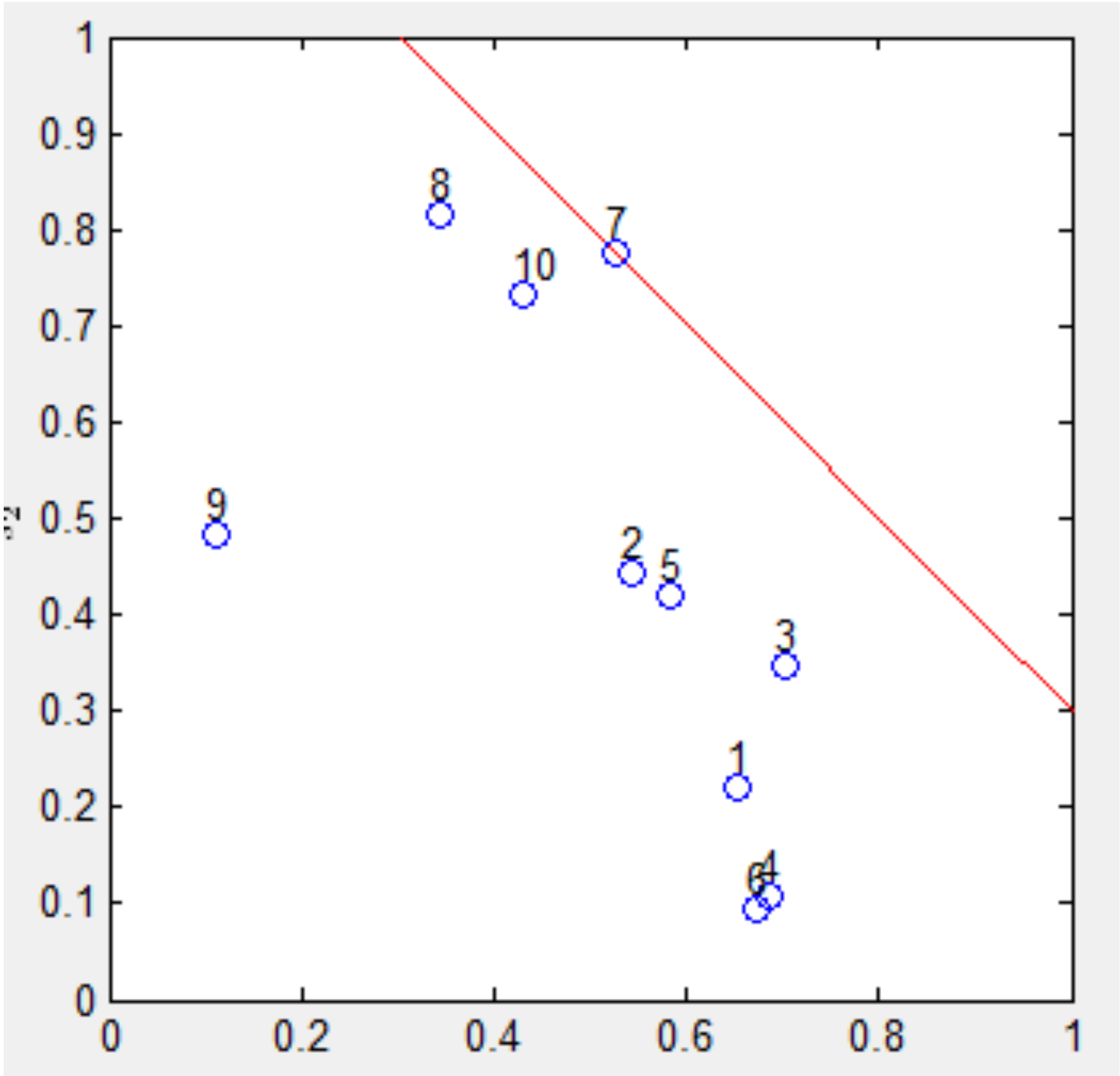
    if $p(F(T_i)>F(T_j))>.5$ and $p(F(T_j)>F(T_k))>.5$
    then $p(F(T_i)>F(T_k))>.5$

  - Besides, $p(F(T_i)>F(T_j))>.5$ iff $s_{i,1} + s_{i,2} > s_{j,1} + s_{j,2}$
  - Complexity: O(N log N)
  - NP-hard for d>2 (weak stochastic transitivity fails)
- ORA under footrule
  - $O(N^{2.5})$ for d=2
    - Min. cost perfect matching of a weighted bipartite graph
  - $O(N^{2^{(d-1)}})$ for d>2
- NB: ORA-footrule is a 2-approximation of ORA-Kendall

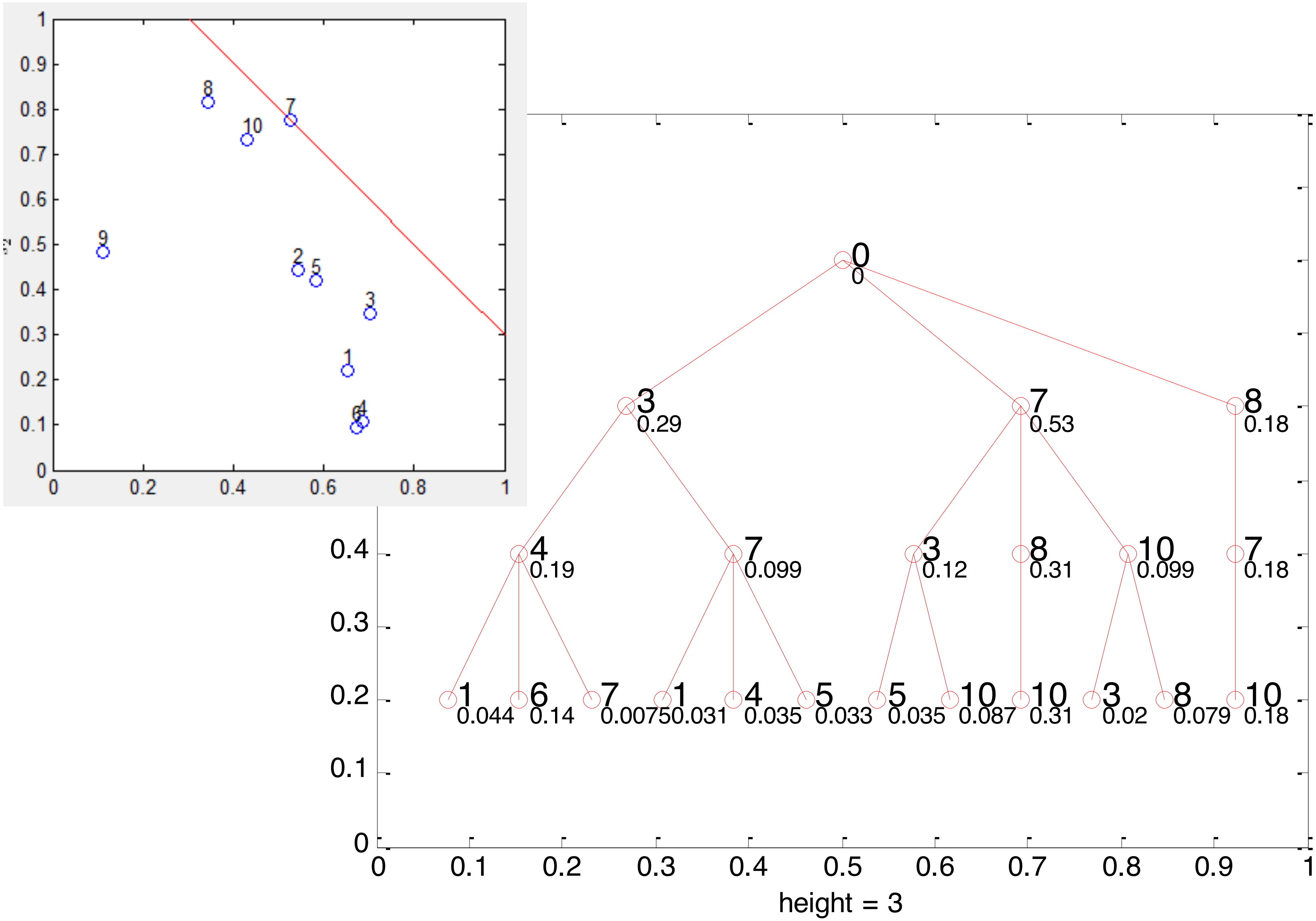
Incremental approach

- Based on an incremental construction of a tree representing the possible orderings
- Each path from the root to a node at depth K represents a possible prefix of length K
- Probability values are assigned to each node
    - (probability of the corresponding prefix)

Points corresponding to join results for d=2

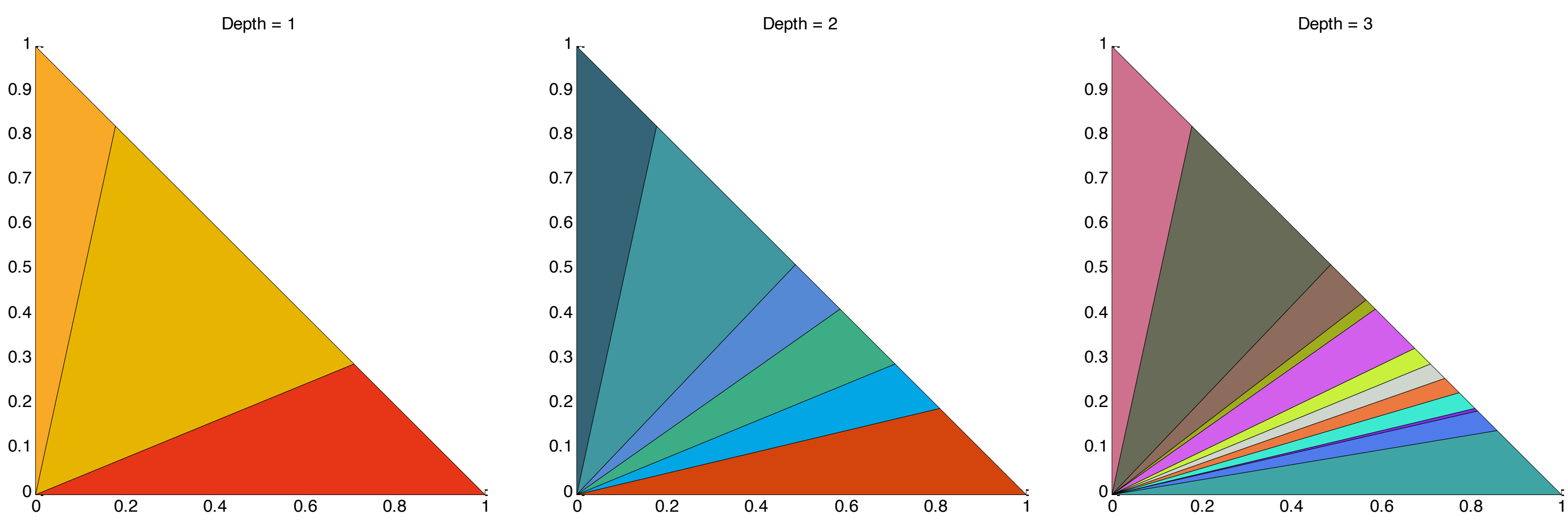


Tree construction

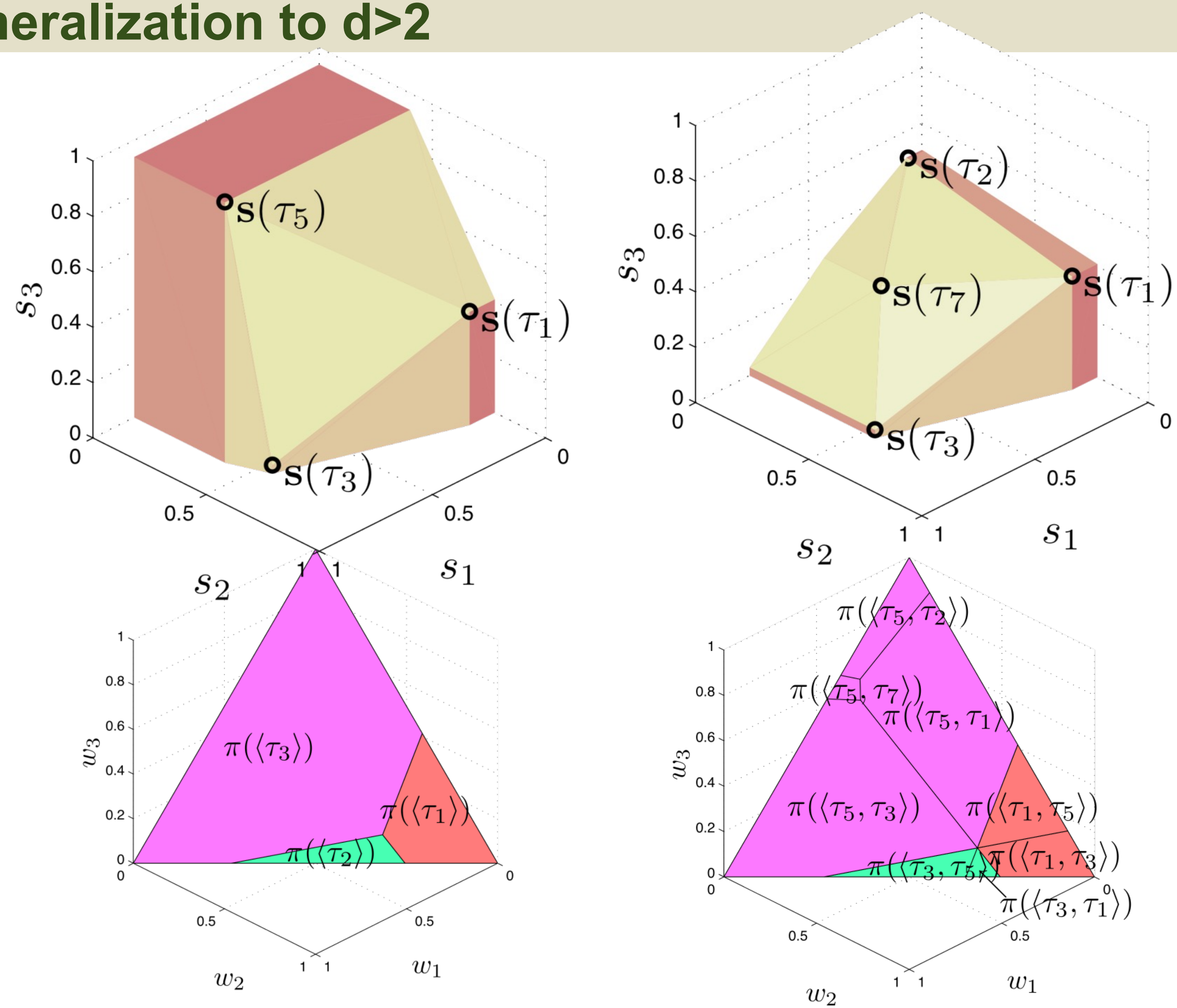


Observations

- Tree generation analyzes only the convex hull of the N points
    - #points on the convex hull $O((\log N)^{d-1})$
- At each level, the regions in the weight space corresponding to a node are polyhedra

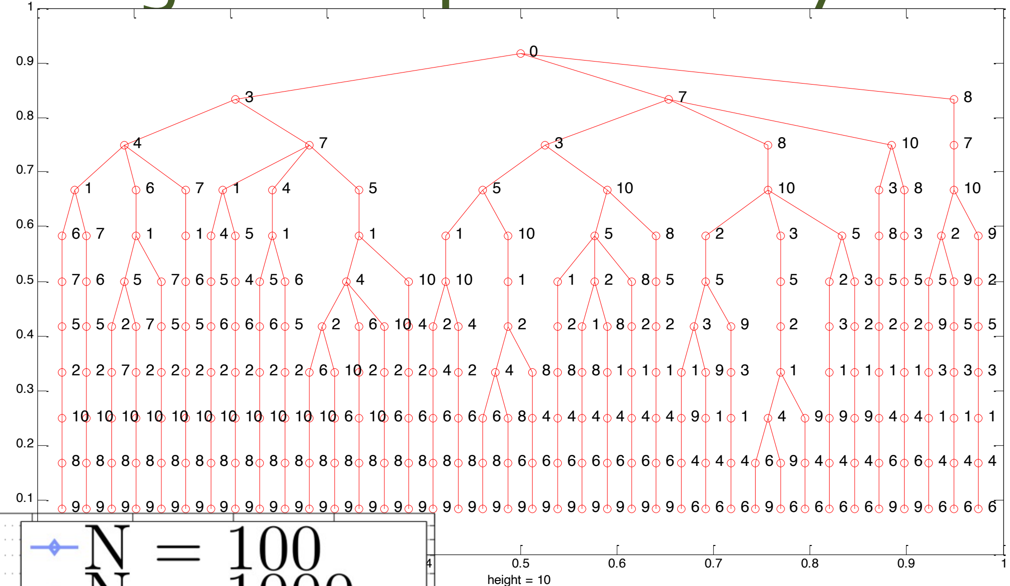


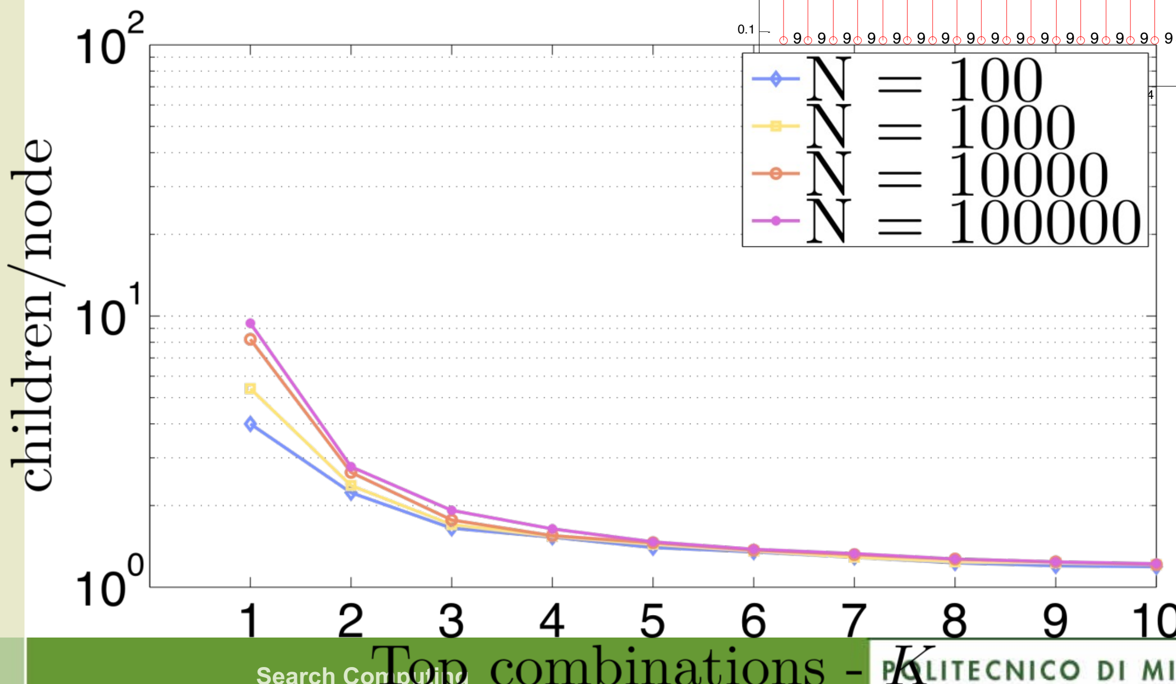
Generalization to d>2



Tree size

- Tree size does not grow exponentially

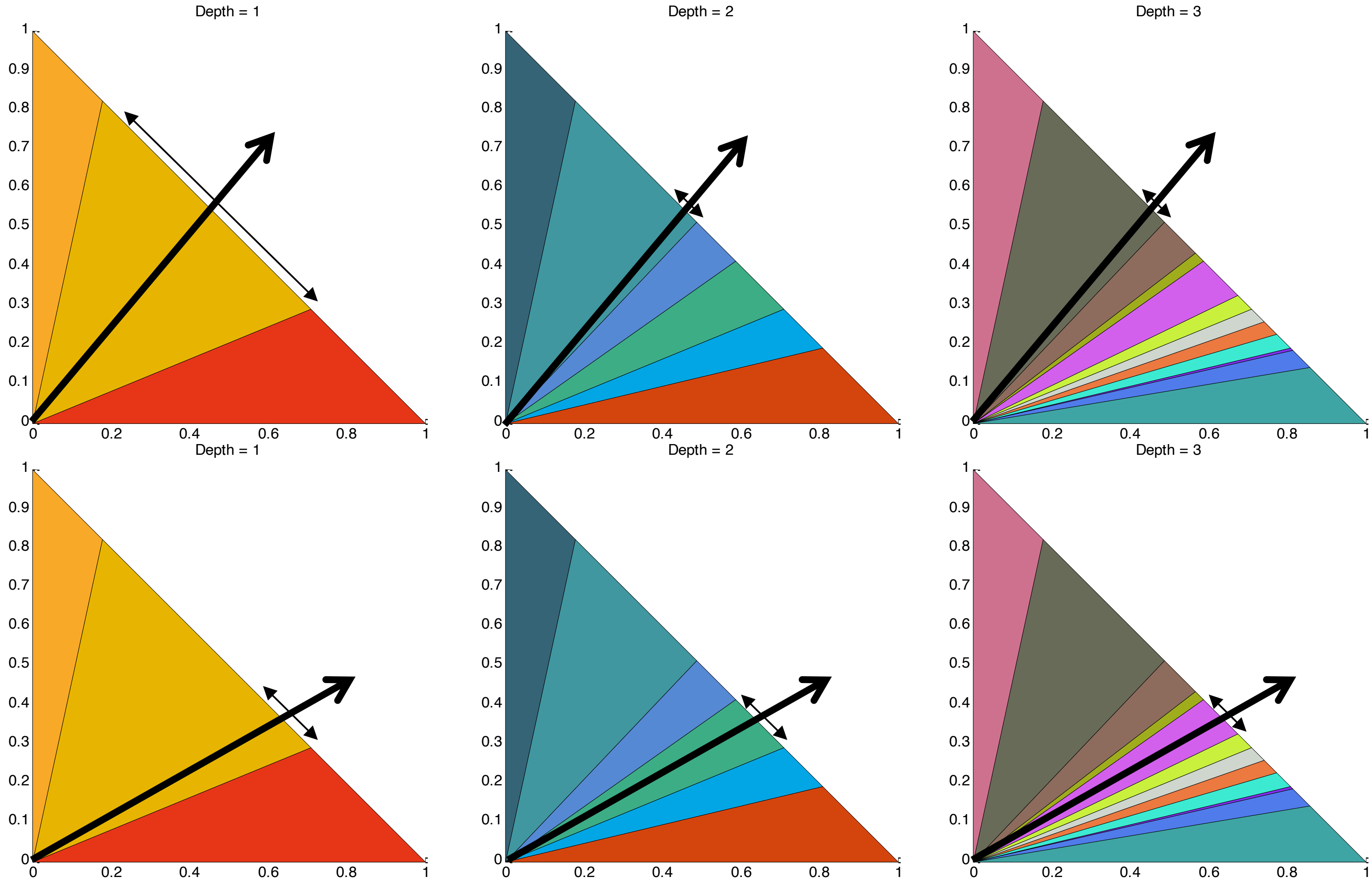




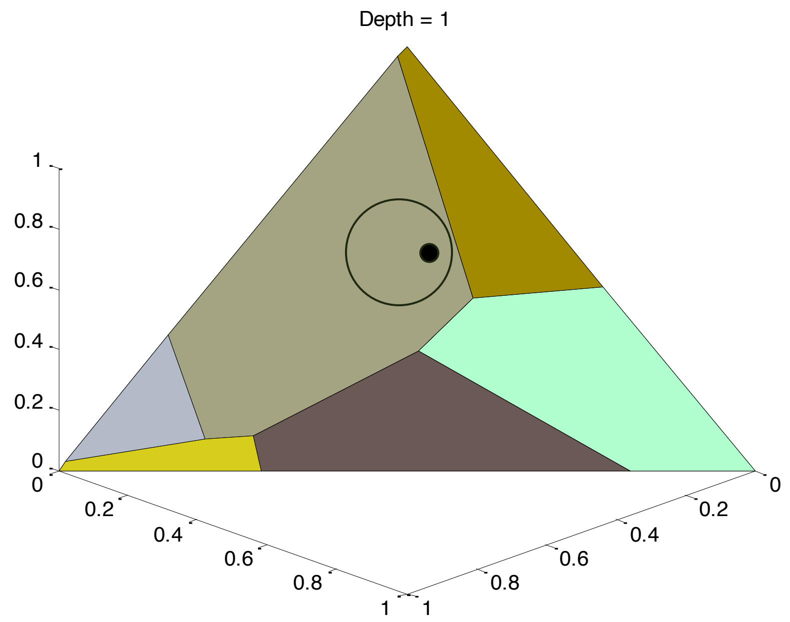
Node probabilities and MPO

- Computing node probabilities amounts to computing volumes of convex polyhedra
  - Shoelace formula…
- This is NP-hard, and thus too expensive, in higher dimensions
- A Monte-Carlo sampling approach is therefore adopted for d>2 (approximate solutions)
- Optimization when searching for MPO:
  - prune branches rooted at a node with probability less than the current MPO candidate

Stability for d=2



Stability for d=3



Summary of problems and complexities

Problem	$\boldsymbol{d=2}$	$\boldsymbol{d=3}$	$\boldsymbol{d>3}$
MPO (average case)	$O(N(logN)^{K+1})$	$O(N(logN)^{2K+1})$	$O(N^{\lfloor d/2 \rfloor+1}(logN)^{(d-1)K})$ [§]
MPO (worst case)	$O(N^2 logN)$	$O(N^4)$	$O(N^{2^{d-1}})$ [§]
ORA (Kendall tau)	$O(NlogN)$	NP-Hard	NP-Hard
ORA (Footrule)	$O(N^{2.5})$	$O(N^4)$	$O(N^{2^{d-1}})$ [§]
STB	$O(N)$	$O(N)$	$O(dN)$
LIK	$O(N)$	$O(N^2)$	$O(N^{2^{d-2}})$ [§]

[§] Approximate solution.

Not discussed in this talk

- Pruning dominated join results
- Preferences among weights
- Experiments

Main References

Historical papers

- Jean-Charles de Borda
  Mémoire sur les élections au scrutin. Histoire de l'Académie Royale des Sciences, Paris 1781
- Nicolas de Condorcet
  Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix, 1785
- Kenneth J. Arrow
  A Difficulty in the Concept of Social Welfare. Journal of Political Economy. 58 (4): 328–346, 1950

Rank aggregation and ranking queries

- Ronald Fagin, Ravi Kumar, D. Sivakumar
  Efficient similarity search and classification via rank aggregation. SIGMOD Conference 2003: 301-312
- Ronald Fagin
  Combining Fuzzy Information from Multiple Systems. PODS 1996: 216-226
- Ronald Fagin
  Fuzzy Queries in Multimedia Database Systems. PODS 1998: 1-10
- Ronald Fagin, Amnon Lotem, Moni Naor
  Optimal Aggregation Algorithms for Middleware. PODS 2001

Skylines and k-Skybands

- Stephan Börzsönyi, Donald Kossmann, Konrad Stocker
  The Skyline Operator. ICDE 2001: 421-430
- Jan Chomicki, Parke Godfrey, Jarek Gryz, Dongming Liang
  Skyline with Presorting. ICDE 2003: 717-719
- Dimitris Papadias, Yufei Tao, Greg Fu, Bernhard Seeger
  Progressive skyline computation in database systems. ACM Trans. Database Syst. 30(1): 41-82 (2005)

Main References

Extensions of skylines: flexible skylines

- Paolo Ciaccia, Davide Martinenghi
  Reconciling Skyline and Ranking Queries. PVLDB 10(11): 1454-1465 (2017)
- Paolo Ciaccia, Davide Martinenghi
  FA + TA < FSA: Flexible Score Aggregation. CIKM 2018: 57-66

Extensions of ranking queries: uncertainty, proximity, diversity

- Mohamed A. Soliman, Ihab F. Ilyas, Davide Martinenghi, Marco Tagliasacchi
  Ranking with uncertain scoring functions: semantics and sensitivity measures. SIGMOD Conference 2011: 805-816
- Davide Martinenghi, Marco Tagliasacchi
  Proximity Rank Join. PVLDB 3(1): 352-363 (2010)
- Piero Fraternali, Davide Martinenghi, Marco Tagliasacchi
  Top-k bounded diversification. SIGMOD Conference 2012: 421-432
- Akrivi Vlachou, Christos Doulkeridis, Yannis Kotidis, Kjetil Nørvåg
  Reverse top-k queries. ICDE 2010: 365-376
- Davide Martinenghi, Marco Tagliasacchi:
  Cost-Aware Rank Join with Random and Sorted Access. IEEE Trans. Knowl. Data Eng. 24(12): 2143-2155 (2012)
- Davide Martinenghi, Marco Tagliasacchi:
  Proximity measures for rank join. ACM Trans. Database Syst. 37(1): 2:1-2:46 (2012)
- Piero Fraternali, Davide Martinenghi, Marco Tagliasacchi:
  Efficient Diversification of Top-k Queries over Bounded Regions. SEBD 2012: 139-146
- Ilio Catallo, Eleonora Ciceri, Piero Fraternali, Davide Martinenghi, Marco Tagliasacchi:
  Top-k diversity queries over bounded regions. ACM Trans. Database Syst. 38(2): 10 (2013)

Main References

Web Access

- Daniele Braga, Stefano Ceri, Florian Daniel, Davide Martinenghi:
  Optimization of multi-domain queries on the web. Proc. VLDB Endow. 1(1): 562-573 (2008)
- Andrea Calì, Davide Martinenghi:
  Conjunctive Query Containment under Access Limitations. ER 2008: 326-340
- Andrea Calì, Davide Martinenghi:
  Querying Data under Access Limitations. ICDE 2008: 50-59
- Andrea Calì, Diego Calvanese, Davide Martinenghi:
  Dynamic Query Optimization under Access Limitations and Dependencies. J. Univers. Comput. Sci. 15(1): 33-62 (2009)
- Andrea Calì, Davide Martinenghi:
  Optimizing Query Processing for the Hidden Web. APWeb 2010: 397
- Andrea Calì, Davide Martinenghi:
  Querying the deep web. EDBT 2010: 724-727