BDTLN + ROOT ERL 7002

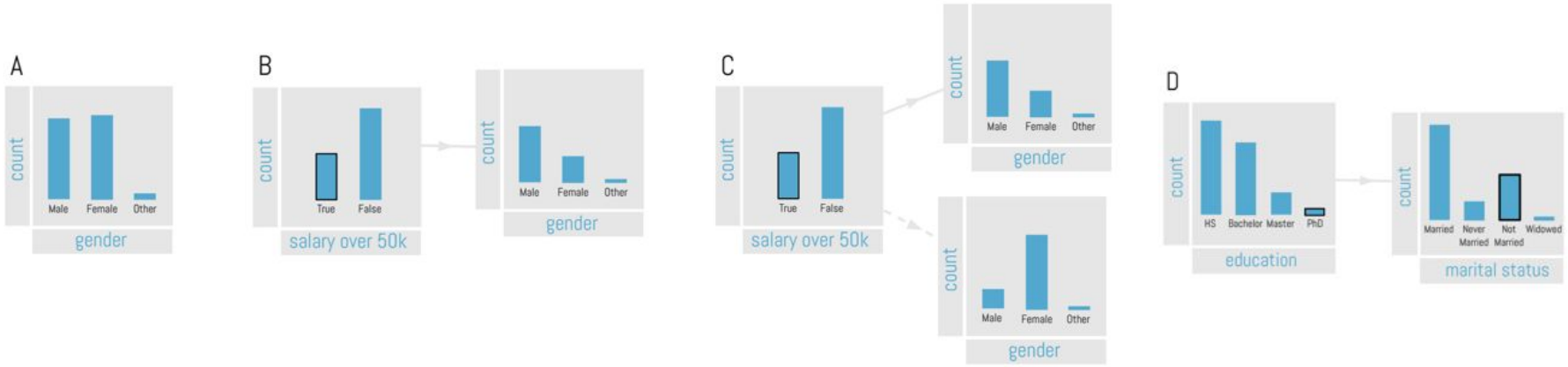
The Travelling Analyst Problem

Orienteering applied to exploratory data analysis

Alexandre Chanson - Nicolas Labroche - Patrick Marcel - Vincent T'Kindt

Exploratory Data Analysis

- Interactive analysis of large volumes of data



- Outcome: a sequence of meaningful queries

- **Challenge** *: how to produce this sequence automatically?*

Comparison Queries

```
select t1.continentexp as "Continent", April as "April", May as "May"
from
(select month, continentexp, sum(cases) as April
from covid
where month = '4'
group by month, continentexp) t1,
(select month, continentexp, sum(cases) as May
from covid
where month = '5'
group by month, continentexp) t2
where t1.continentexp = t2.continentexp;
```

sum of cases

Continent	April	May
Europe	863874	608110
Oceania	2812	467
America	1104862	1404912
Africa	31598	92626
Asia	333821	537584

$$\pi_{a,val,val'}((\gamma_{a,agg(m)\rightarrow val}(\sigma_{b=val}(R))) \bowtie (\gamma_{a,agg(m)\rightarrow val'}(\sigma_{b=val'}(R))))$$

Comparison is a common activity for data analysts (Zgraggen 2018)

Comparison queries - Limit the search space

continentexp	countriesandterritories	day	month	cases	deaths
America	Curacao	29	5	1	0
America	Curacao	28	5	0	0
America	Curacao	27	5	0	0
America	Curacao	26	5	1	0
America	Curacao	25	5	1	0
America	Curacao	24	5	0	0
America	Curacao	23	5	0	0
America	Curacao	22	5	0	0
America	Curacao	21	5	0	0
America	Curacao	20	5	0	0
America	Curacao	19	5	0	0
America	Curacao	18	5	0	0
America	Curacao	17	5	0	0
America	Curacao	16	5	0	0

Schema : *A* categoricals attributes, *M* measures, *G* aggregations {min, max, sum, avg}

Instance (10^3 lines)

Active Domain : *v* values

$$\pi_{a,val,val'}((\gamma_{a,agg(m)\rightarrow val}(\sigma_{b=val}(R))) \bowtie (\gamma_{a,agg(m)\rightarrow val'}(\sigma_{b=val'}(R))))$$

$$\text{Search space} = (\sum_{i \in [1 \dots n]} {}_{adom(Ai)}C_2) \cdot (|A|-1) \cdot |G| \cdot |M|$$

Problem statement

Given all possible comparison queries on a database :

- Find the sequence of queries
- Maximizing interestingness
- Such that it can be executed in limited time budget
- The distance over the sequence is minimal.

- The distance can be turned into an epsilon-bound.
- The execution time is given by DBMS

Distance between queries

- **Metric adapted from [Aligon 14]**

Q1

Age Group	Average deaths	
	India	UK
0-18	4	2
18-25	43	12
25-50	641	607
50-65	3000	3777
65+	5600	4286

Intuitively, d(Q1, Q2) < d(Q1, Q3)

Q2

Age Group	Average deaths	
	Germany	UK
0-18	4	2
18-25	43	12
25-50	641	607
50-65	3000	3777
65+	5600	4286

Q3

Eye color	Average deaths	
	Germany	UK
Blue	23	34
Green	43	45
Brown	567	345

Interest of a query

- **Interest = p-value of statistical tests validating the query result as evidence of an insight (Zgraggen 18)**

Age Group	Average deaths	
	India	UK
0-18	4	2
18-25	43	12
25-50	641	607
50-65	3000	3777
65+	5600	4286

- Insight -> H0“Deaths are independent of Age”
- One evidence: average deaths in UK and India by Age Group
- Verification -> Statistical test
- Interest -> 1 - p-value

Problem Formulation

- Extension of the Orienteering Problem, with a service time
- Model based on formulations from (Kara 16, Gunawan 16)

Data

$c_{i,j}$ the distance between queries, q_i and q_j. (Positive integer)
t_i denotes the execution time of q_i (Positive integer)
v_i is theinterestingness score associated with q_i (real number in [0,1])

Variables

$x_{i,j}, (i,j) \in 1..n, x_{i,j} = 1$ if q_i comes directly before q_j in the solution, 0 otherwise
$x_{0,i}, i \in 1..n, x_{i,j} = 1$ if q_i is the first query of the solution, 0 otherwise
$x_{i,n+1}, i \in 1..n, x_{i,j} = 1$ if q_i is the last query of the solution, 0 otherwise
$s_i, i \in 1..n$: boolean variables denoting the presence of q_i in the solution.
$u_i, i \in 1..n$: integer variables used in subtour elimination constraints.

Model

objective

$$max \sum_{i=1}^{n} v_i s_i \quad (1)$$

under constraints

$$\sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} c_{i,j} x_{i,j} \leq \epsilon_d \quad (2)$$

$$\sum_{i=1}^{n} t_i s_i \leq \epsilon_t \quad (3)$$

$$\sum_{i=0, j\neq i}^{n} (x_{i,j}) - s_j = 0, \forall j \in 1..n \quad (4)$$

$$\sum_{j=1, j\neq i}^{n+1} (x_{i,j}) - s_i = 0, \forall i \in 1..n \quad (5)$$

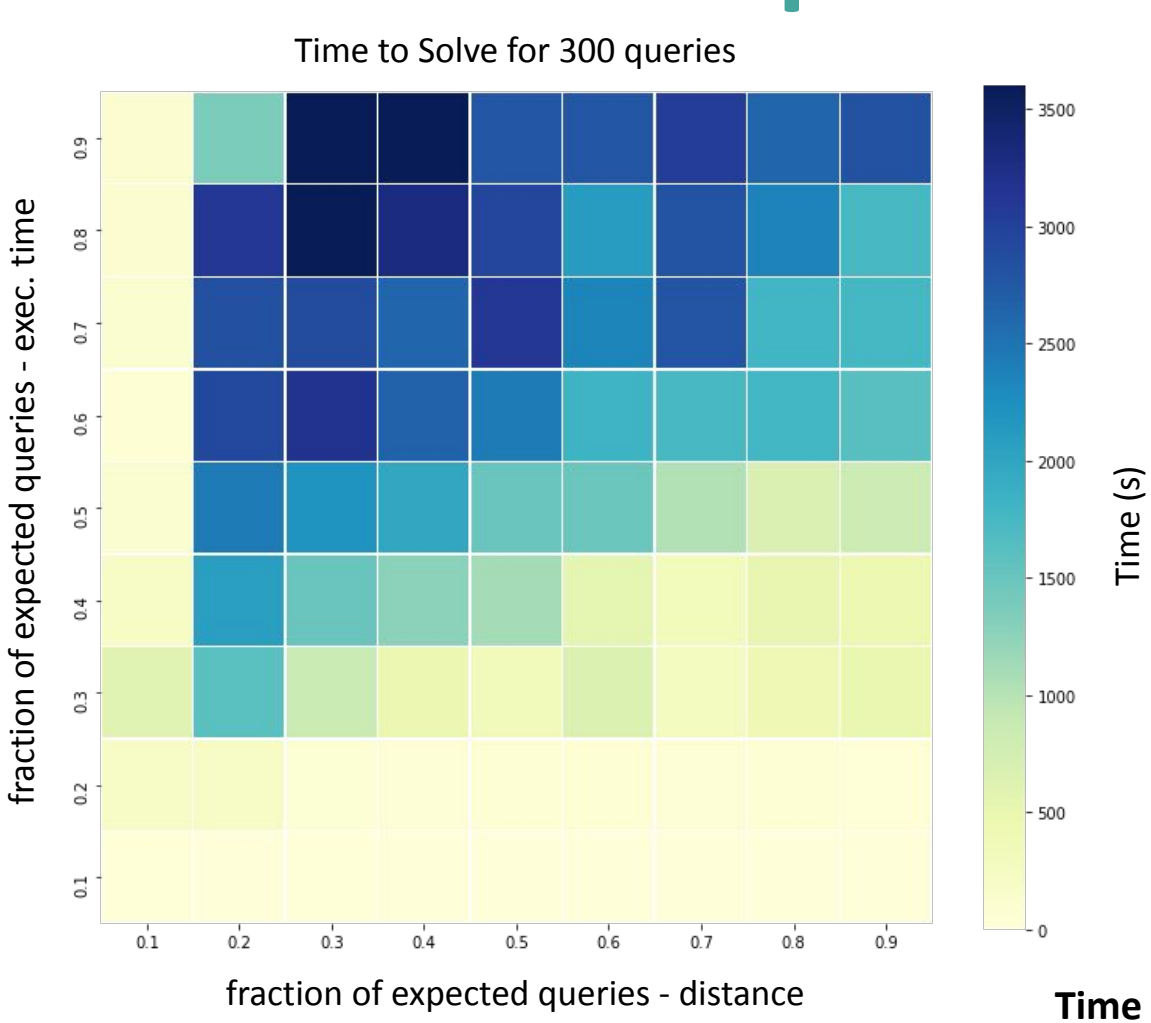
$$\sum_{j=1}^{n} x_{0j} = \sum_{i=1}^{n} x_{i,n+1} = 1 \quad (6)$$

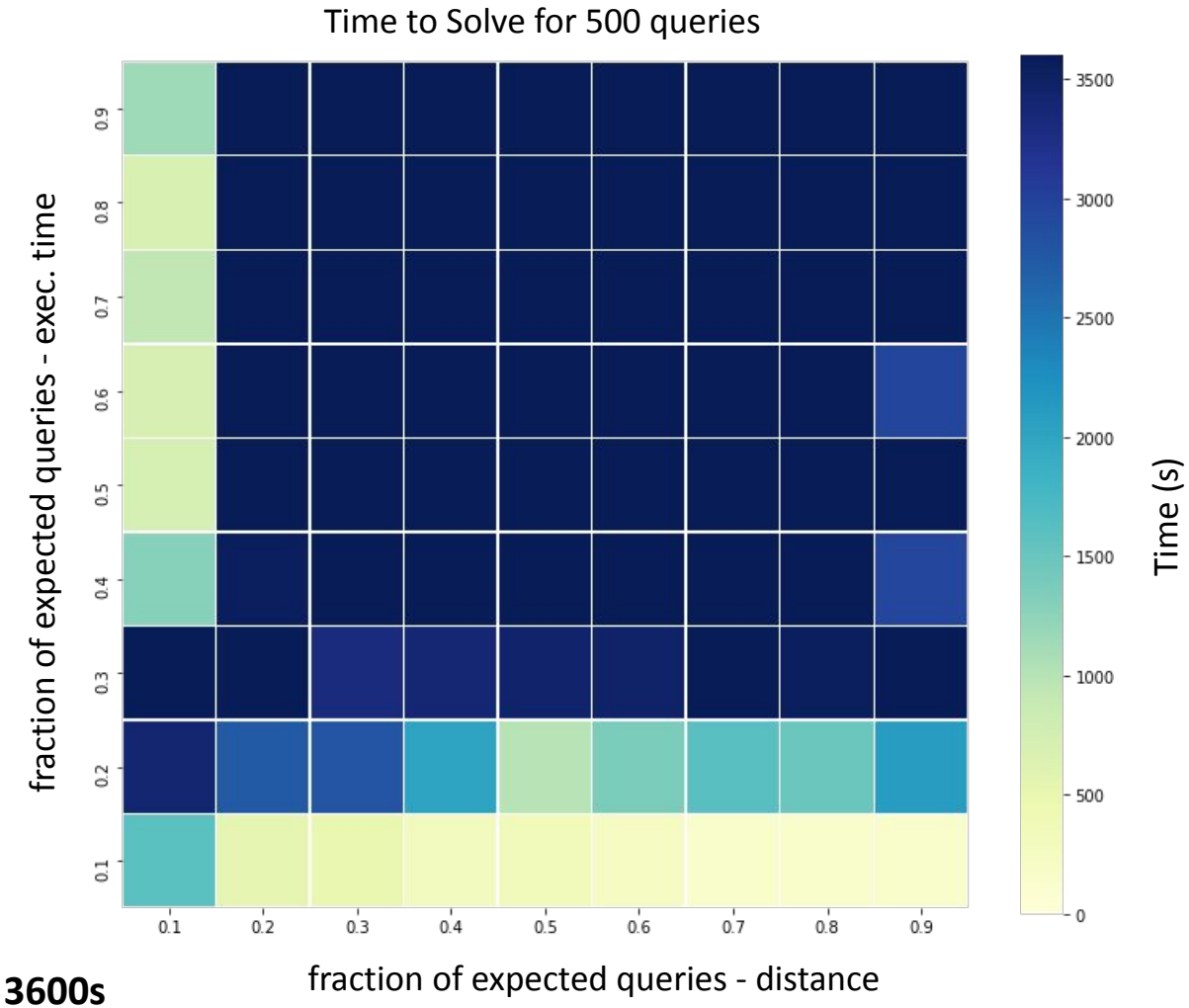
$$2 \leq u_i \leq n, i \in 1..n \ , \ u_i - u_j + 1 \leq (n-1)(1 - x_{ij}), (i,j) \in 1..n \quad (7)$$

Influence of epsilon-constraints

- Distances (integer) drawn from uniform distribution U(1,10)
- Interestingness (real) drawn from U(0,1)
- time (integer) drawn from U(5,50)
- Epsilon constraints expressed as a fraction of expected queries based on distribution mean

Influence of epsilon-constraints





Time OUT : 3600s

Constraint generation strategy

- Constraint generation strategy for subtour elimination (Pferschy 17)
  - Preliminary experiments shows substantial speedups on hard instances

Perspectives

- Exact solution is still tractable for smallest real instances (> 10^3 queries)
- Mathheuristics (VPLS) for large instance (> 10^4 queries)
- ‘Faster’ Heuristics for larger instances of the problem (> 10^6 queries)
- Addressing more generic query pattern (> 10^9 queries)
- Extension of the problem where a limited disk space can be used (indexing) to speed up some queries

References

Aligon 14 : Similarity measures for OLAP sessions, Julien Aligon, Matteo Golfarelli, Patrick Marcel, Stefano Rizzi, Elisa Turricchia, Knowledge and information systems 2014

Kara 16: New formulations for the orienteering problem, Imdat Kara, Papatya Sevgin Bicakci, and Tusan Derya, Procedia Economics and Finance, 39 :849–854, 2016

Zgraggen 18: Investigating the Effect of the Multiple Comparisons Problem in Visual Analysis, Emanuel Zgraggen, Zheguang Zhao, Robert Zeleznik, and Tim Kraska, In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems

Gunawan 16: Orienteering problem : A survey of recent variants, solution approaches and applications, A. Gunawan, H. C. Lau, and P. Vansteenwegen, E.J.O.R., 255(2), 2016

Pferschy 17: Generating subtour elimination constraints for the TSP from pure integer solutions, Ulrich Pferschy & Rostislav Staněk, Central European Journal of Operations Research 2017

BDTLN + ROOT ERL 7002

Thank You for Listening

The Travelling Analyst Problem

Alexandre Chanson - chanson@univ-tours.fr - github.com/AlexChanson