A Model for Astronomical Cross-Matching Disambiguation

Vinícius P. Freire (UFC)
Fábio Porto (LNCC)
José A. F. de Macêdo (UFC)

DEXL LAB
EXTREME DATA LAB

arida
Advanced Research In Database

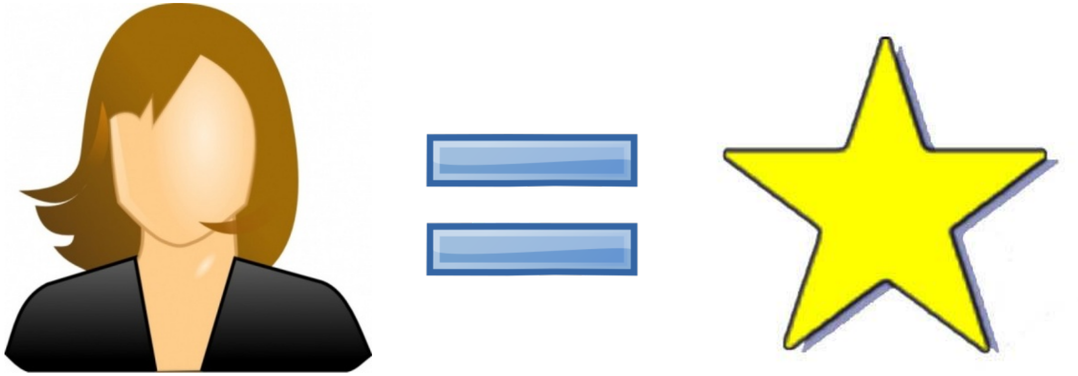
LIneA
Laboratório Interinstitucional de e-Astronomia

HOSCAR Meeting
Bordeaux - France

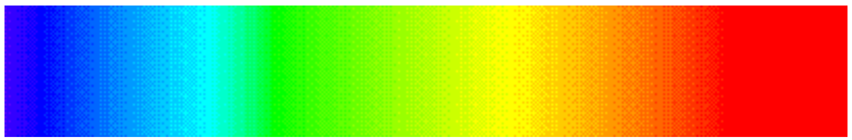
Agenda

- Introduction
  - Fundamentals
    - Indexing structures
    - Algorithms
    - Experiments and evaluation
  - Motivation and goal
- A disambiguation model proposal
- Conclusion

What is a catalog?



Spectroscopic Survey :



Introduction

Different Astronomical Surveys (Catalogs)

Sloan Digital Sky Survey
Mapping the Universe

DARK ENERGY SURVEY

2MASS
2 MICRON ALL SKY SURVEY

La Sagra Sky Survey
LSSS
MPC J75

Southern Sky Survey

Introduction

- Surveys produce catalogs with intersections in the covered area of the sky;
- Problem:
  - Getting an integrated view provided by different catalogs requires data cross-matching
  - How to identify celestial objects that appear in different catalogs with descriptive variations?

Introduction

- Problem identified as "Entity Resolution"
  - Identify instances of objects from different databases that match the same real world entity
- Alternatives for entity resolution in the “cross-matching catalogs“ problem:
  - use the position of the objects in the sky (coordinate system based on RA, DEC);
  - use other attributes to help treating the ambiguities.

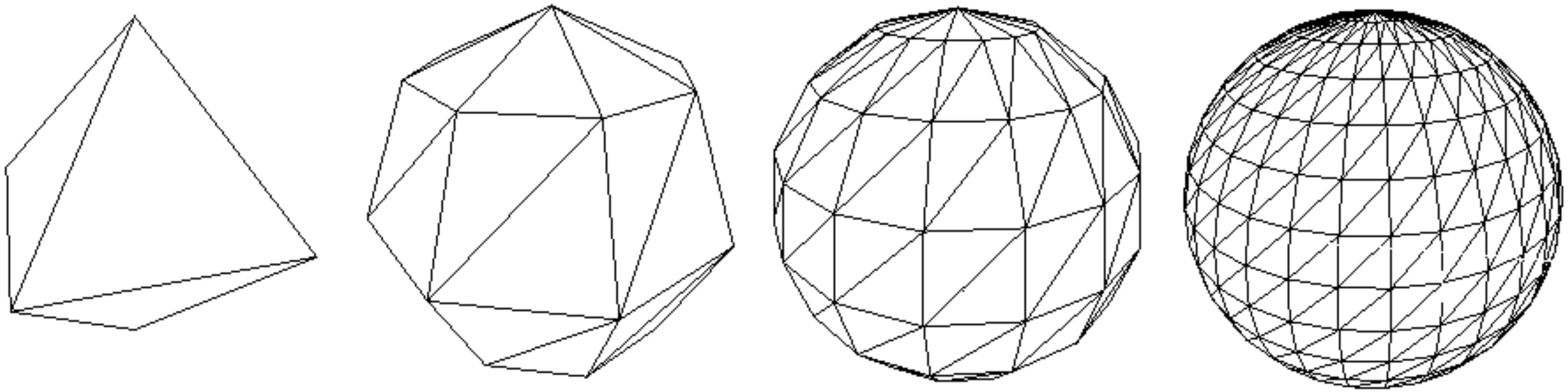
DEXL LAB EXTREME DATA LAB

Current solutions

- Support cross-matching by location
  - Main Strategy: represent the sky making the use of data structures in order to facilitate the localization of the stars and their neighbors in space
  - use the spatial indexing structures to support cross-matching:
    - HTM (Hierarchical Triangular Mesh)
    - Q3C (Quad Tree Cube)
    - Zones

Fundamentals

HTM
(Hierarchical Triangular Mesh)

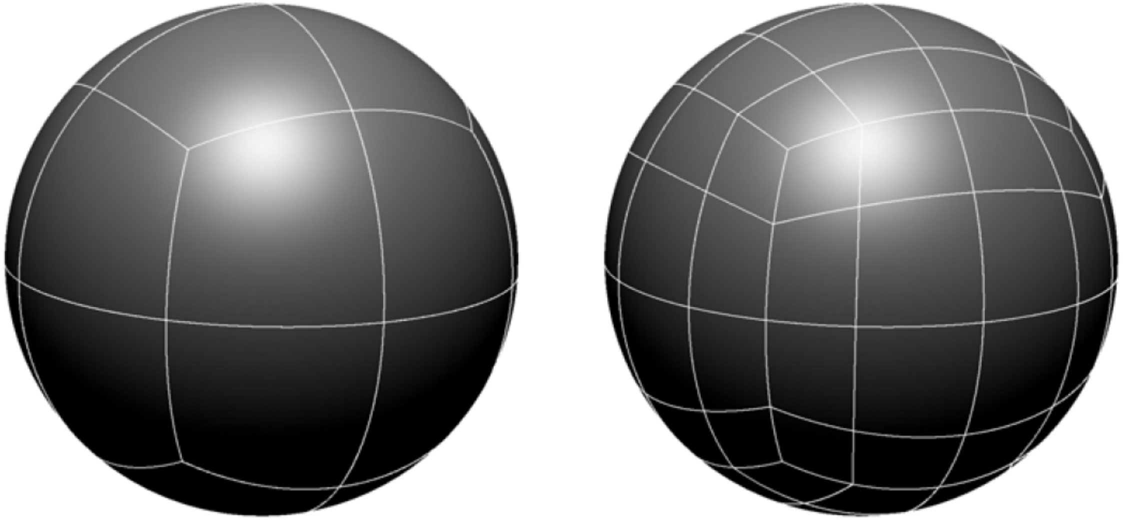


Kunszt, P. Z., Szalay, A. S., and Thakar, A. R. (2001). The hierarchical triangular mesh.

DEXL LAB
EXTREME DATA LAB

Fundamentals

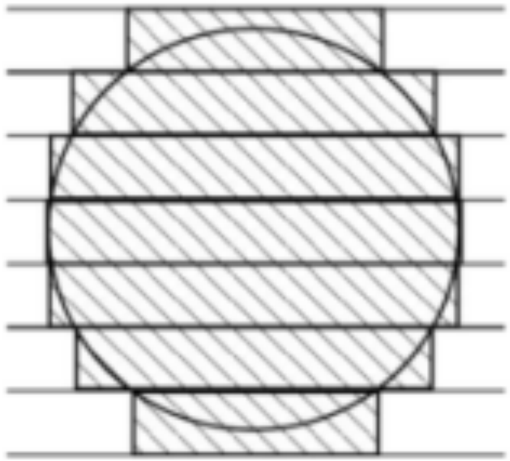
Q3C
(Quad Tree Cube)



Koposov, S. and Bartunov, O. (2006). Q3C , Quad Tree Cube – The new Sky-indexing Concept for Huge Astronomical Catalogues and its Realization for Main Astronomical

Fundamentals

ZONES



[Gray, J., Szalay, A. S., Thakar, A. R., and et al. (2004). There goes the neighborhood: Relational algebra for spatial data search.]

Fundamentals

Binary cross-matching of catalogs

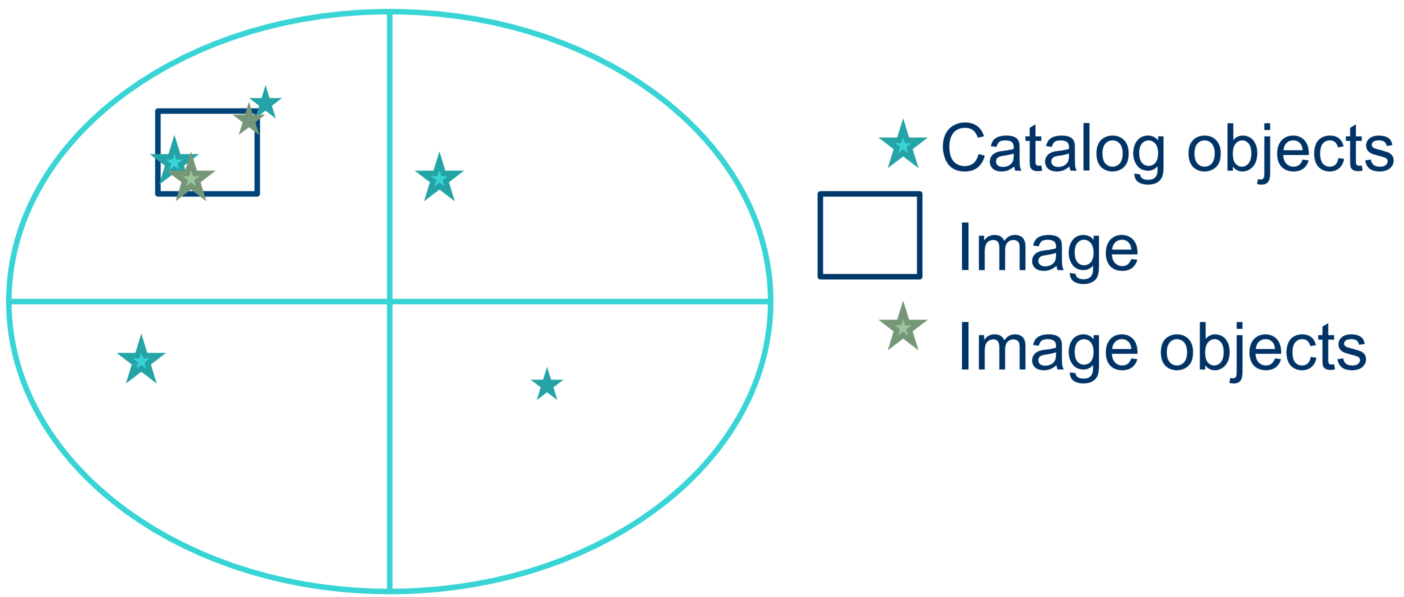
- Support the cross-matching by location

- Algorithms
  - Fast Approximate matching
  - Q3c Join

Fundamentals

Algorithms

[Fu et al. 2012] **Fast approximate matching of astronomical objects.**



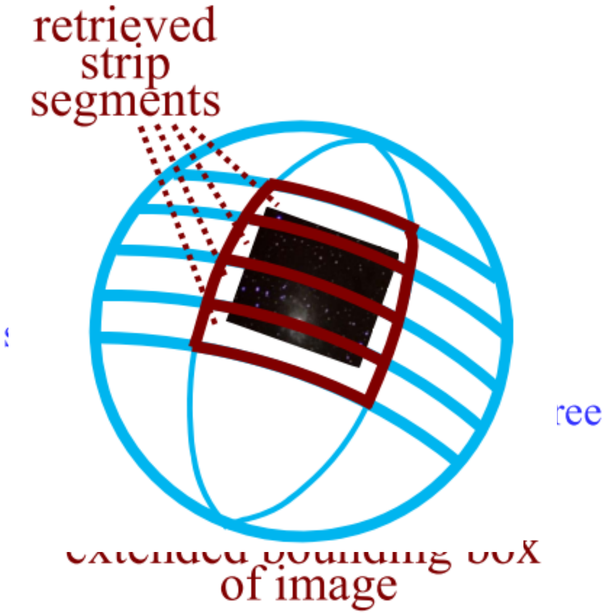
DEXL LAB EXTREME DATA LAB

Fundamentals

Matching process

1. Index by strips

2. Given an image, retrieve the catalog objects that are similar to the image objects;



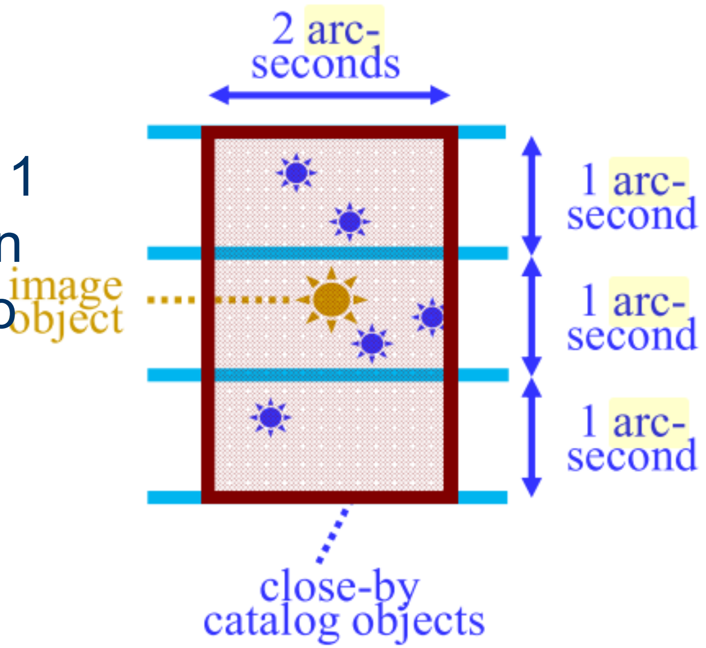
3. A matching catalog object must be within 1 arcsec from the image object
   - Extend the image bounding box of 1 arcsec in order to find all objects.

4. Retrieve in memory all the catalog objects contained in the area of the extended image

Fundamentals

Matching process (cont.)

5. Recovered strips are divided into thinner sub-strips (1 arc-second)
   - For each image object:
     - Retrieve the catalog objects that are at 1 arcsec distant, considering only their own substrips, as well as the ones in their two adjacent substrips.
     - The condition to achieve the match is:
       - For each image object **p**, it is assumed that its nearest catalog object is **q**, and the **q** nearest object is also **p**. Then **q** is the match for **p**.



DEXL LAB EXTREME DATA LAB

Fundamentals

Some considerations about the Fast approximate cross-matching

- Positive points
  - Indexing relatively fast
    - Index a catalog of 2 billion objects in less than two hours
  - Fast matching
    - Match a catalog of 2 billion objects and an image of 100,000 objects in 4 seconds.
- Negative point
  - Matching using more than 2 catalogs does not generate symmetric results

DEXL LAB EXTREME DATA LAB

Fundamentals

Cross-Matching using Q3C

- Implementation in PostgreSQL (q3c_join())

```
Select *
from table1, table2
where q3c_join(table1.ra, table1.dec, table2.ra, table2.dec,
   0.001);
```

- Supposing there is a Q3C index over table2
  - function q3c_join():
    - Defines 4 range queries to approximate the crossmatch circles:
      - If the object of table1 is within these ranges, then the matching is achieved.

DEXL LAB EXTREME DATA LAB

Fundamentals

Experiments using Q3C

- Goal: To evaluate the quality of binary cross-matching based on a spatial criterion
- Test environment:
  - Catalogs Involved:
    - 2MASS (470,992.70 objects)
    - BCC v.05 (1,376,582,713 objects)
  - Radius 0.001 degree
- Result
  - Matching of 17,701,306 objects
  - Processing: 142 seconds

DEXL LAB EXTREME DATA LAB

Fundamentals

Evaluating the matching

- Difficulty - No prior knowledge
- Test environment :
  - Catalogs Involved:
    - 2MASS (470,992,970 objects)
    - 2MASS (470,992,970 objects)
  - Radius 0.001 degree
- Result:
  - The obvious would return 470,992,970 elements
  - Returned 483,197,616 objects
  - 12,204,646 (2.6% of the total) were near objects with different positions
  - 6,102,323 (1.3% of the total) were the real number of false positives - two equal tables
- Why did it happen?

DEXL LAB
EXTREME DATA LAB

Fundamentals

Ambiguity

- Although these mistakes represent just 1.3% of the total, the amount of ambiguous matchings was very high (millions of objects).

- Matching ambiguity is an open problem and needs to be explored.

Motivation

- Ambiguity
  - Binary matching does not generate symmetric results using more than 2 catalogs
  - There are no solutions to n-way matching
  - The best attribute which identifies the astronomical objects is its position, but it isn't precise
- All these characteristics produce ambiguities

Goal

- Measure this ambiguity and propose a better solution to ambiguous n-way matching

Model proposal

- Use a probabilistic model

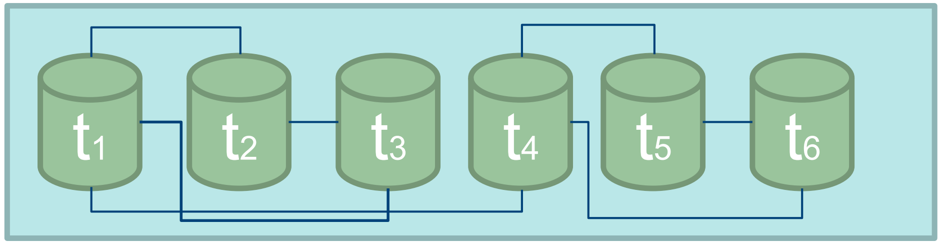
  - Associate a probability distribution for possible matchings

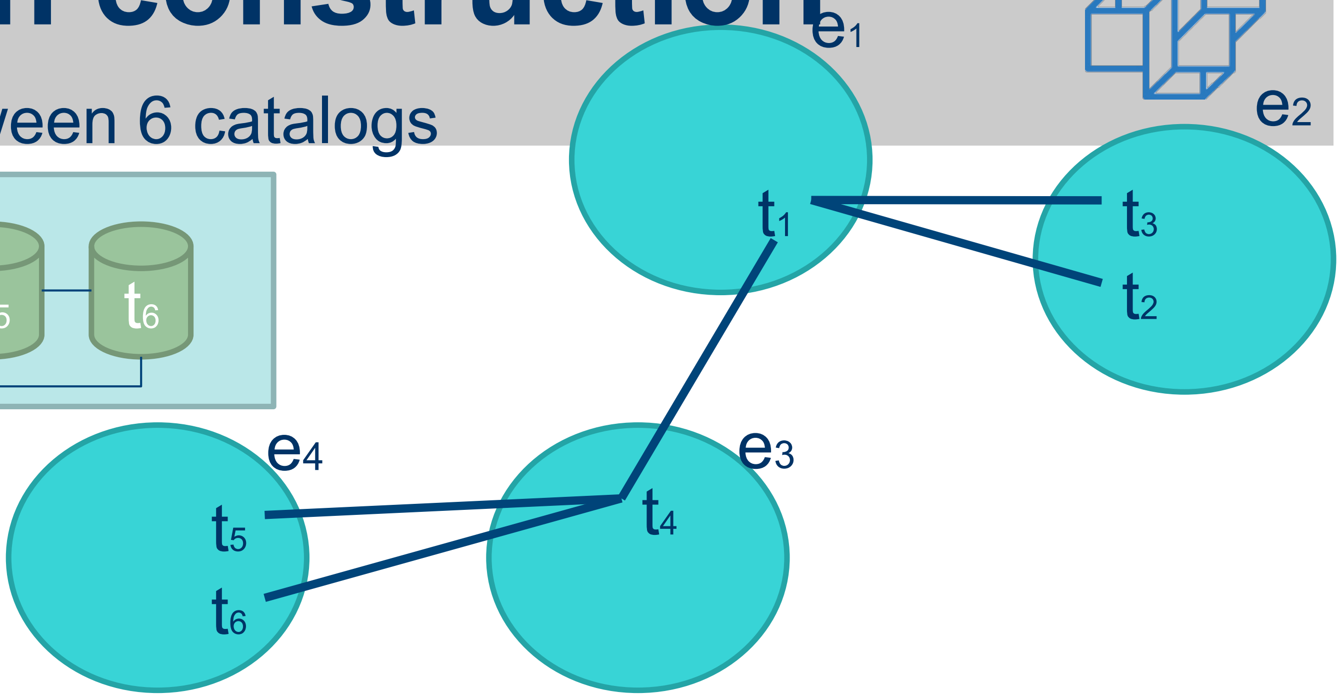
  - Produce all possible worlds and calculate its probability based on a model proposed in [Ayat, N., Akbarinia, R., Afsarmanesh, H., Valduriez, P. Entity Resolution for Uncertain Data. (2012)]

DEXL LAB
EXTREME DATA LAB

Model in construction

Matching between 6 catalogs





Initial matching algorithm classifies as 4 different objects. Each cluster represents an entity.

Ge_i	**Candidates objects to be e_i**
Ge_1	t_1, t_2, t_3, t_4
Ge_2	t_1, t_2, t_3
Ge_3	t_1, t_4, t_5, t_6
Ge_4	t_4, t_5, t_6

Different Possible Worlds

Ambiguous matching

$W_1 = \{e_1 = \{t_1, t_2, t_3\}, e_2 = \{\}, e_3 = \{t_4\}, e_4 = \{t_5, t_6\}\}$

$W_2 = \{e_1 = \{t_1, t_2\}, e_2 = \{t_3\}, e_3 = \{t_4\}, e_4 = \{t_5, t_6\}\}$

$W_3 = \{e_1 = \{t_1\}, e_2 = \{t_2, t_3\}, e_3 = \{t_4, t_5\}, e_4 = \{t_6\}\}$

Main directives

- Problems:
  - Probabilistic model to calculate the probability of each world;
  - Efficient algorithm for choosing the best world
- Expectations:
  - Generate a n-way more precise matching algorithm
  - Solve ambiguities;
- How to evaluate the quality of the result?

Model in construction

Object	Entity	Probability
t_1	e_1	0.2
t_1	e_2	0.3
t_1	e_3	0.5
t_2	e_1	0.7
t_2	e_2	0.3
t_3	e_1	0.4
t_3	e_2	0.6
t_4	e_1	0.2
t_4	e_3	0.3
t_4	e_4	0.5
t_5	e_3	0.3
t_5	e_4	0.7
t_6	e_3	0.6
t_6	e_4	0.4

Probability of the object t_i represent the entity e_j

$W_1 = \{e_1 = \{t_1, t_2, t_3\}, e_2 = \{\}, e_3 = \{t_4\}, e_4 = \{t_5, t_6\}\}$

$$P(W_i) = \begin{cases} \prod_{i=1}^{n} P(e_i) & \text{se } e_i \in W \\ 1 & \text{se } e_i \notin W \end{cases}$$

$$P(e_i) = \begin{cases} \prod_{t \mid t \in Ge_i} P(t, e_i) & \text{se } t \in e_i \\ 1 - P(t, e_i) & \text{se } t \notin e_i \end{cases}$$

$P(W_1) = 0.0448 \times 1 \times 0.042 \times 0.14 = 0.0003$

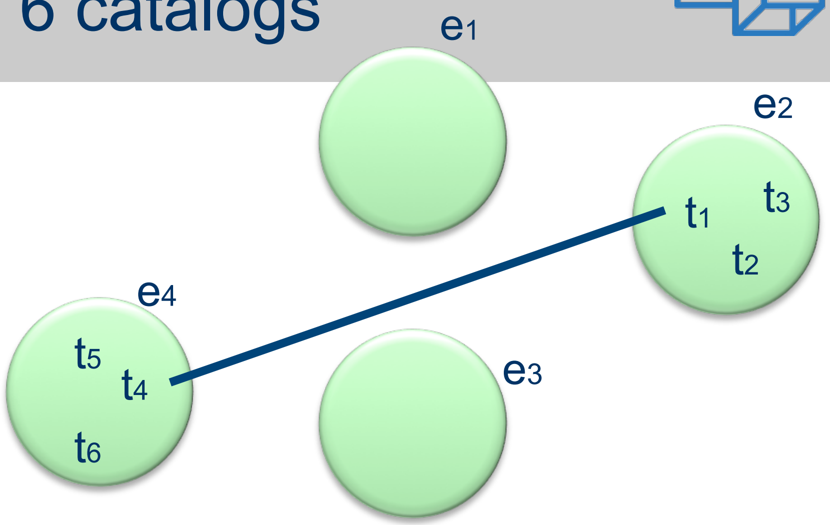
e_i	$P(e_i)$
e_1	0.2 x 0.7 x 0.4 x (1 - 0.2) = 0.0448
e_2	1
e_3	0.3 x (1 - 0.5) x (1 - 0.3) x (1 - 0.6) = 0.042
e_4	0.7 x 0.4 x (1 - 0.5) = 0.14

Ge_i	**Candidates objects to be e_i**
Ge_1	t_1, t_2, t_3, t_4
Ge_2	t_1, t_2, t_3
Ge_3	t_1, t_4, t_5, t_6
Ge_4	t_4, t_5, t_6

DEXL LAB EXTREME DATA LAB

Model in construction

Matching between 6 catalogs



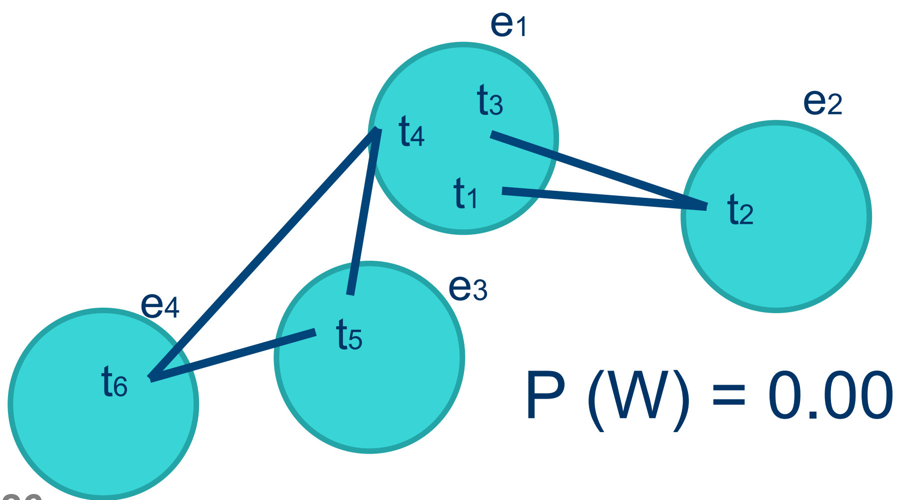
P(W) = 0.00756

144 POSSIBLE WORLDS

$$P(W) = \begin{cases} \prod_{i=1}^{n} P(e_i) & \text{se } e_i \in W \\ 1 & \text{se } e_i \notin W \end{cases}$$

$$P(e_i) = \begin{cases} \prod_{t|t \in Ge_i} P(t, e_i) & \text{se } t \in e_i \\ 1 - P(t, e_i) & \text{se } t \notin e_i \end{cases}$$

Ge_i	**Candidates objects to be** e_i
Ge_1	t_1, t_2, t_3, t_4
Ge_2	t_1, t_2, t_3
Ge_3	t_1, t_4, t_5, t_6
Ge_4	t_4, t_5, t_6



P (W) = 0.000001

DEXL LAB
EXTREME DATA LAB

Conclusion

- To develop this work, it is necessary:
  - Decide which initial algorithm should be used
  - Find a way to calculate the probability of an object belonging to an entity
  - Define the best probabilistic model to calculate the probability of each world;
  - Develop an efficient algorithm for choosing the best world
  - Find a way to evaluate the quality of the result

Merci

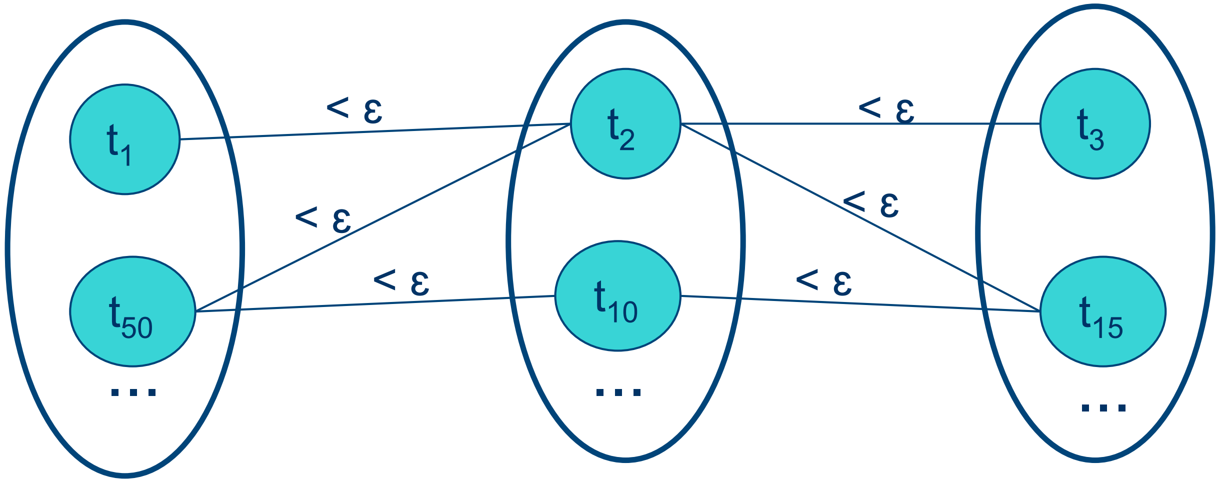
Vinícius P. Freire (UFC/LNCC)
vinipires@lia.ufc.br

Another way to view this problem

Hypergraph

Hypernode = Catalog

Edge means possibility of matching



Fundamentals

[Fu et al. 2012] Fast approximate matching of astronomical objects.

- For each object ***p*** in the image, we are looking for an object ***q*** in the catalog such that:
  - Among all objects in the image, ***p*** is the nearest to **q**.
  - Among all objects in the catalog, ***q*** is the nearest to ***p***.
  - The distance between ***p*** and ***q*** in the two-dimensional spherical coordinates is at most 1 arc second, which is 1/3600 of a degree.