NACluster: A Non-Supervised Clustering Algorithm for Matching Multi Catalogues

(UFC)
Fábio Porto (LNCC)
José A. F. de Macêdo (UFC)
Reza Akbarinia (INRIA)

Advanced Research In Database

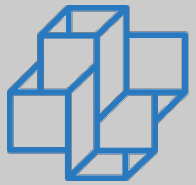
DEXL LAB
EXTREME DATA LAB

Fourth Brazil-France Workshop

On High Performance Computing and Scientific Data Management Driven by Highly Demanding Applications

LIneA

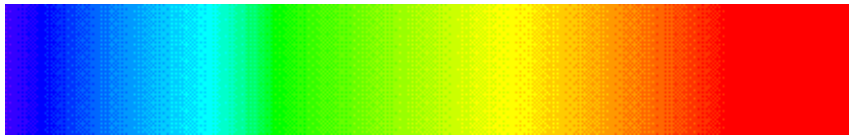
Agenda



- Introduction
  - Fundamentals
  - Motivation and goal
- NACluster Algorithm
  - Experiments
- Future Works

What is an astronomical catalog?

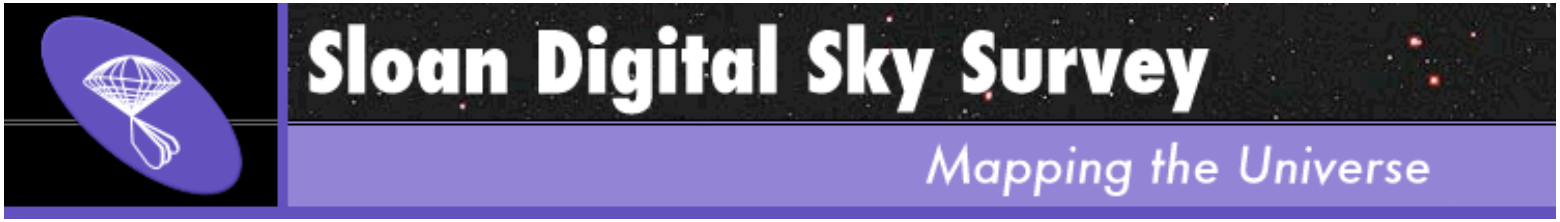
Spectroscopic Survey :

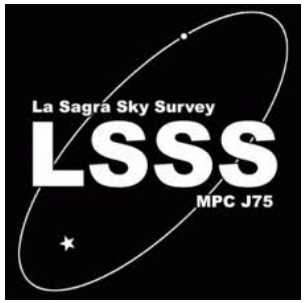


DEXL LAB
EXTREME DATA LAB

Introduction

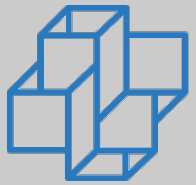
Different Astronomical Surveys (Catalogs)





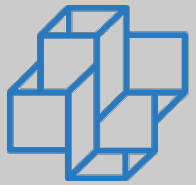
DEXL LAB
EXTREME DATA LAB

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.

Big-Data (in science) Data Challenges

- Data Representation
  - Different Data Models:
    - Data structure and query languages
    - Graphs, Matrixes, Key-Value,…
- **Data Uncertainty**
  - **Data is uncertain**
  - **uncertainty quantification on data**
- Data Partitioning
  - in sync with data processing
- Data Heterogeneity
  - Data Granularity

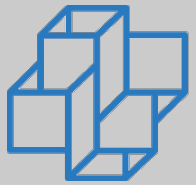
Fundamentals

Current solutions

- Binary cross-matching of catalogs

What can happen if I add a Catalog C to do cross-matching with this result?

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

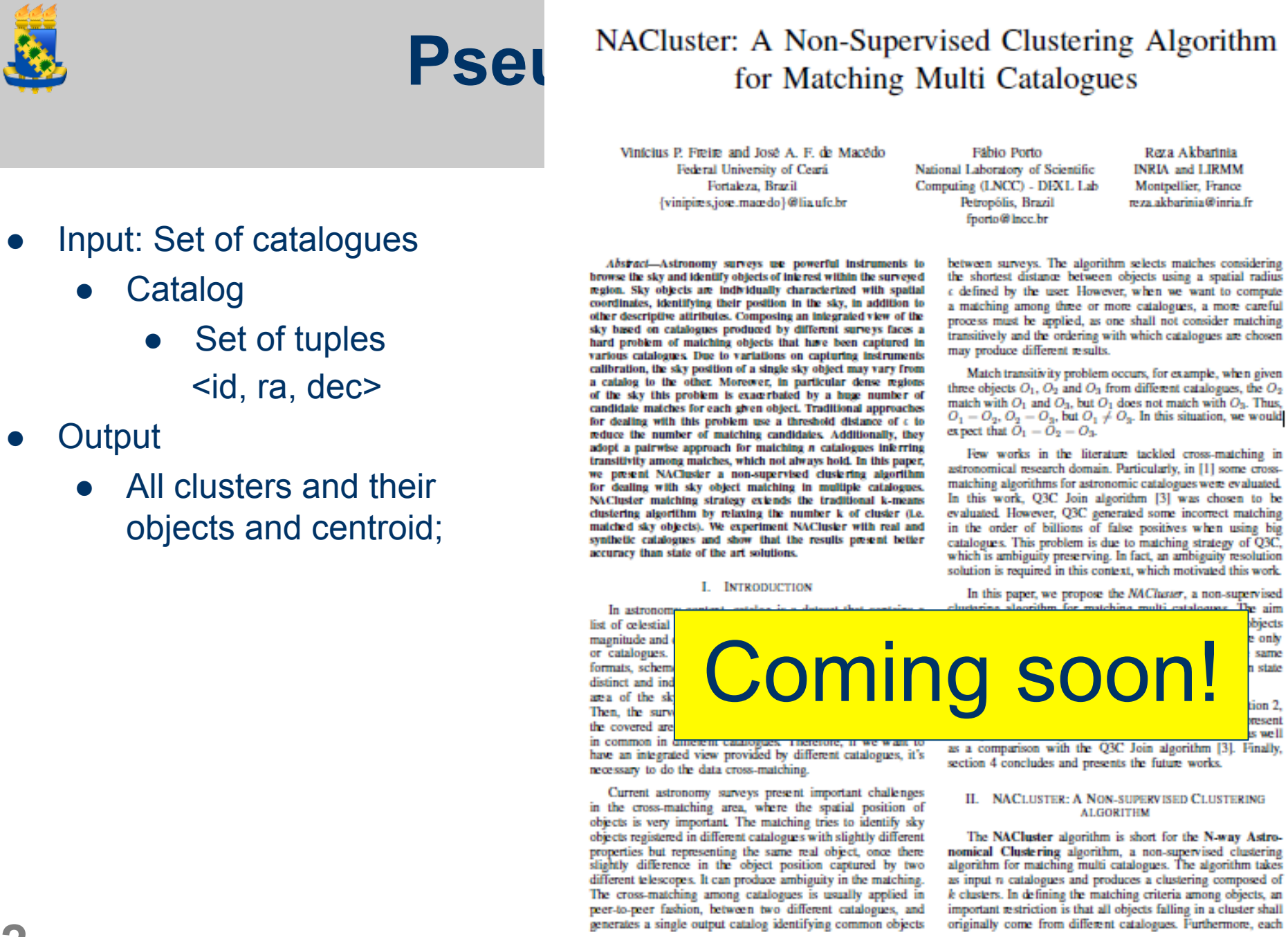
- Propose a solution to treat ambiguous n-way catalog matching

NACluster Algorithm

- **NACluster**
  - **N**-way **A**stronomical **Cluster**ing algorithm
  - Non-supervised clustering algorithm for matching multi catalogues
- Aim
  - Split into clusters the celestial objects present in **N** catalogues
    - In each cluster there are only objects from different catalogues and representing the same real object.
- This contribution allows the improvement on state of the art astronomy catalog matching.

Pseu

- Input: Set of catalogues
  - Catalog
    - Set of tuples
      <id, ra, dec>
- Output
  - All clusters and their objects and centroid;



NACluster: A Non-Supervised Clustering Algorithm for Matching Multi Catalogues

Vinícius P. Freire and José A. F. de Macêdo
Federal University of Ceará
Fortaleza, Brazil
{vinipires,jose.macedo}@lia.ufc.br

Fábio Porto
National Laboratory of Scientific Computing (LNCC) - DEXL Lab
Petrópolis, Brazil
fporto@lncc.br

Reza Akbarinia
INRIA and LIRMM
Montpellier, France
reza.akbarinia@inria.fr

***Abstract*—Astronomy surveys use powerful instruments to browse the sky and identify objects of interest within the surveyed region. Sky objects are individually characterized with spatial coordinates, identifying their position in the sky, in addition to other descriptive attributes. Composing an integrated view of the sky based on catalogues produced by different surveys faces a hard problem of matching objects that have been captured in various catalogues. Due to variations on capturing instruments calibration, the sky position of a single sky object may vary from a catalog to the other. Moreover, in particular dense regions of the sky this problem is exacerbated by a huge number of candidate matches for each given object. Traditional approaches for dealing with this problem use a threshold distance of ϵ to reduce the number of matching candidates. Additionally, they adopt a pairwise approach for matching n catalogues inferring transitivity among matches, which not always hold. In this paper, we present NACluster a non-supervised clustering algorithm for dealing with sky object matching in multiple catalogues. NACluster matching strategy extends the traditional k-means clustering algorithm by relaxing the number k of cluster (i.e. matched sky objects). We experiment NACluster with real and synthetic catalogues and show that the results present better accuracy than state of the art solutions.**

I. Introduction

In astronomy context, catalog is a dataset that contains a list of celestial magnitude and or catalogues. formats, schem distinct and ind area of the sk Then, the surv the covered are in common in different catalogues. Therefore, if we want to have an integrated view provided by different catalogues, it's necessary to do the data cross-matching.

Current astronomy surveys present important challenges in the cross-matching area, where the spatial position of objects is very important. The matching tries to identify sky objects registered in different catalogues with slightly different properties but representing the same real object, once there slightly difference in the object position captured by two different telescopes. It can produce ambiguity in the matching. The cross-matching among catalogues is usually applied in peer-to-peer fashion, between two different catalogues, and generates a single output catalog identifying common objects between surveys. The algorithm selects matches considering the shortest distance between objects using a spatial radius ϵ defined by the user. However, when we want to compute a matching among three or more catalogues, a more careful process must be applied, as one shall not consider matching transitively and the ordering with which catalogues are chosen may produce different results.

Match transitivity problem occurs, for example, when given three objects O_1, O_2 and O_3 from different catalogues, the O_2 match with O_1 and O_3, but O_1 does not match with O_3. Thus, $O_1 - O_2$, $O_2 - O_3$, but $O_1 \neq O_3$. In this situation, we would expect that $O_1 - O_2 - O_3$.

Few works in the literature tackled cross-matching in astronomical research domain. Particularly, in [1] some cross-matching algorithms for astronomic catalogues were evaluated. In this work, Q3C Join algorithm [3] was chosen to be evaluated. However, Q3C generated some incorrect matching in the order of billions of false positives when using big catalogues. This problem is due to matching strategy of Q3C, which is ambiguity preserving. In fact, an ambiguity resolution solution is required in this context, which motivated this work.

In this paper, we propose the *NACluster*, a non-supervised clustering algorithm for matching multi catalogues. The aim objects e only same n state tion 2, resent s well as a comparison with the Q3C Join algorithm [3]. Finally, section 4 concludes and presents the future works.

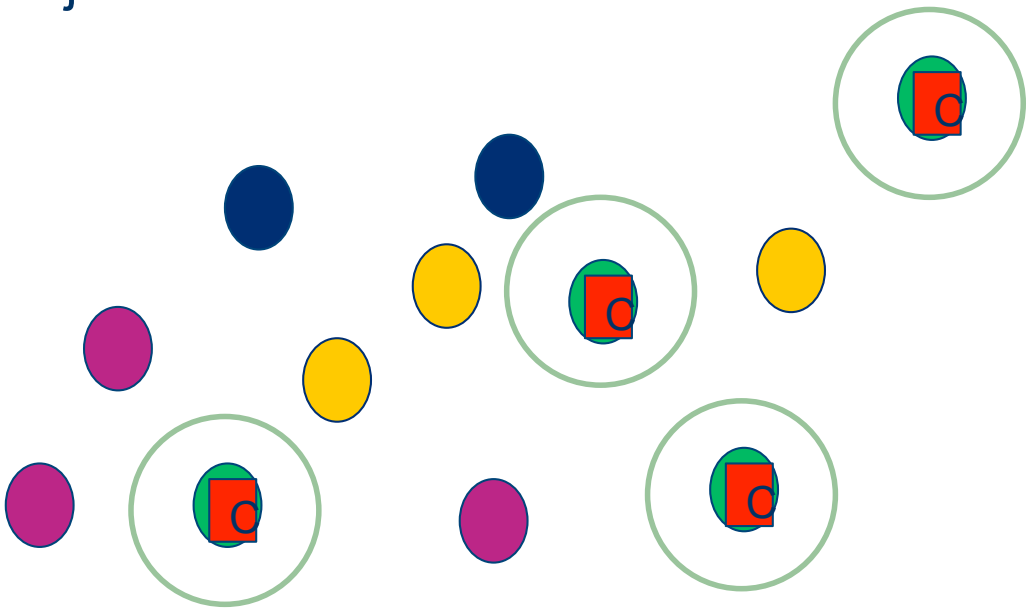
II. NACluster: A Non-Supervised Clustering Algorithm

The **NACluster** algorithm is short for the **N-way Astronomical Clustering** algorithm, a non-supervised clustering algorithm for matching multi catalogues. The algorithm takes as input n catalogues and produces a clustering composed of k clusters. In defining the matching criteria among objects, an important restriction is that all objects falling in a cluster shall originally come from different catalogues. Furthermore, each

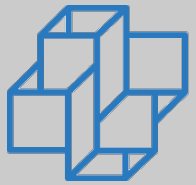
Coming soon!

Pseudocode

- Step 1
  - Initialize Clusters
    - The largest catalog is selected and one cluster is created for each its object.
    - The position of each object is taken as a cluster centroid.

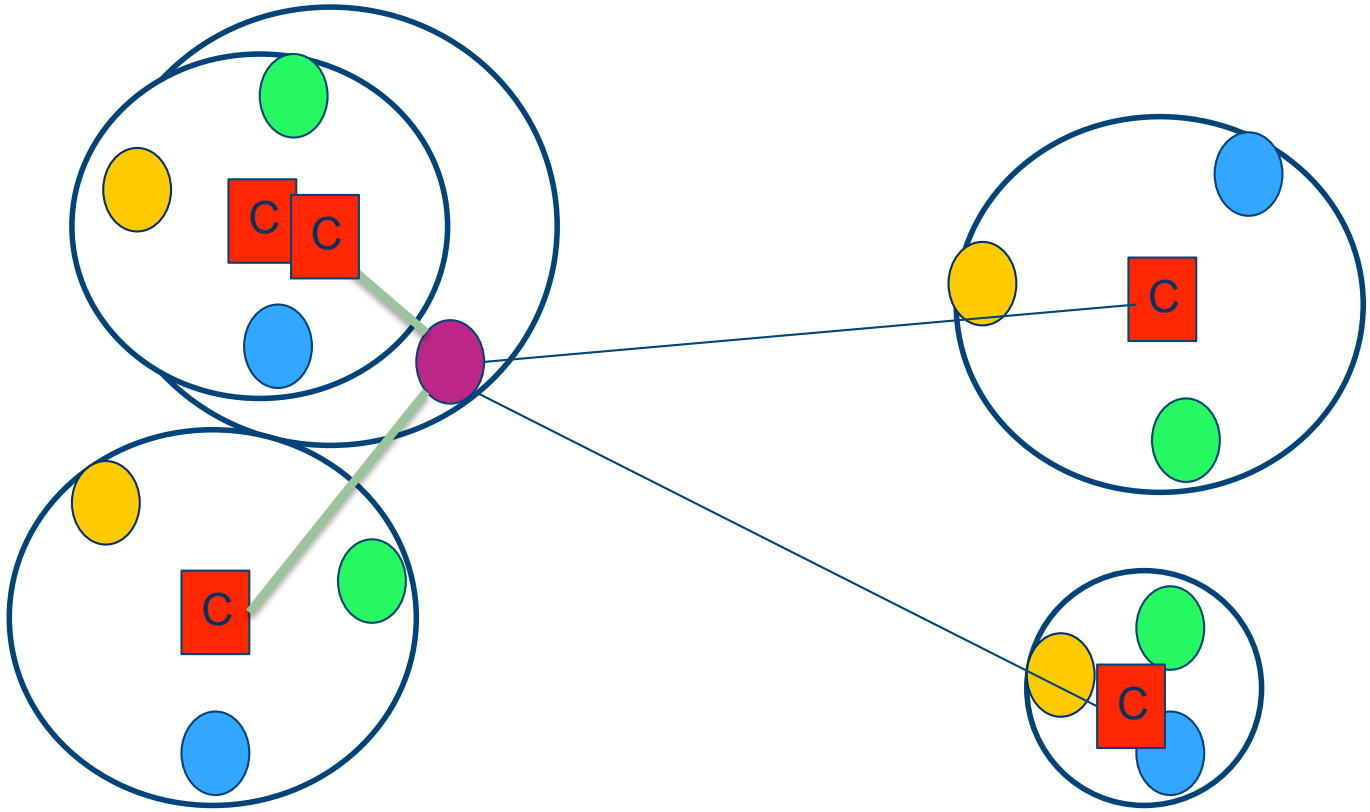


Pseudocode



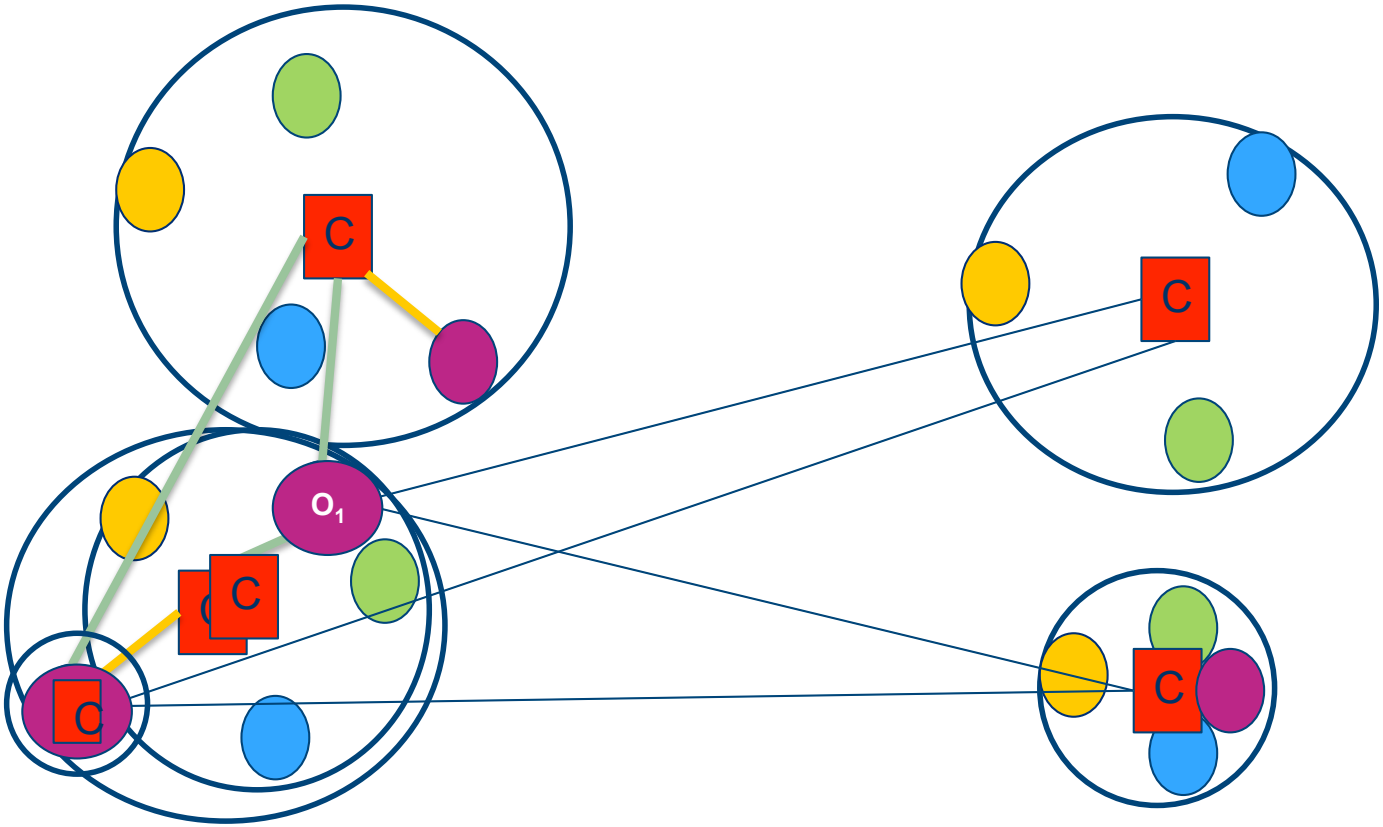
- Step 1
  - Initialize Clusters
    - The largest catalog is selected and one cluster is created for each its object.
    - The position of each object is taken as a cluster centroid.
- The idea of the algorithm is to compare each object of each catalog to all computed cluster centroids, one catalog at a time, by computing the Euclidian distance d(Oi;Ca) of an object Oi to a centroid Ca.

Situation 1



DEXL LAB
EXTREME DATA LAB

Situation 2



DEXL LAB
EXTREME DATA LAB

Pseudocode

- Stop condition
  - After all objects have clustered
    - New Iteration
    - Reset clusters
      - Keeping the centroids positions
  - Finish
    - When the centroids are stable, i.e. all the computed centroids of an iteration are the same as the previous iteration.
- Complexity:
  - Exponential complexity on the number of individual sky elements (i.e. clusters).
  - By using a spatial indexing strategy, the actual number of comparison is reduced to a local region.

Experiments

- Goal: To evaluate the quality of NACluster algorithm
- Test environment:
  - 5 experiments
  - Dataset -> Catalogs Involved:
    - Dense part of 2MASS (90,000 objects)
    - New catalogues generated by a normal distribution function from 2MASS, simulating in this way real variations of the same catalog.

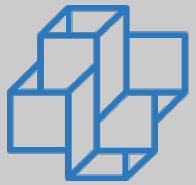
No. of catalogues	Precision	Recall	F-Measure
2	0.9750	0.9763	0.9757
3	0.9717	0.9727	0.9722
4	0.9654	0.9671	0.9662
5	0.9699	0.9713	0.9706
6	0.9734	0.9745	0.9739

Experiments

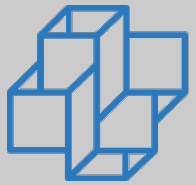
- Test environment:
  - Catalogs Involved:
    - UBVRI Catalog (49,167 Objects)
    - against other 1 to 5 synthetic generated versions of it.

No. of catalogues	Precision	Recall	F-Measure
2	0.9937	0.9938	0.9938
3	0.9944	0.9944	0.9941
4	0.9954	0.9960	0.9957
5	0.9932	0.9932	0.9926
6	0.9913	0.9914	0.9906

Future Works



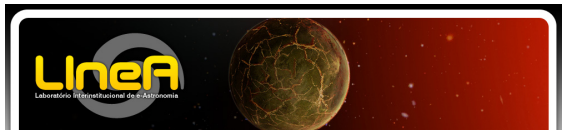
- Parallel Strategy
  - Data Partitioning (see Daniel Gaspar pres.)
  - Spatial indexing strategy in order to reduce the complexity.
- Big Data
  - Catalogs – 1 Billion Objects



Obrigado

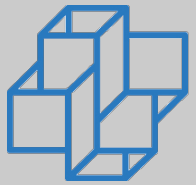
Vinícius P. Freire
vinipires@lia.ufc.br

arida
Advanced Research In Database



DEXL LAB
EXTREME DATA LAB

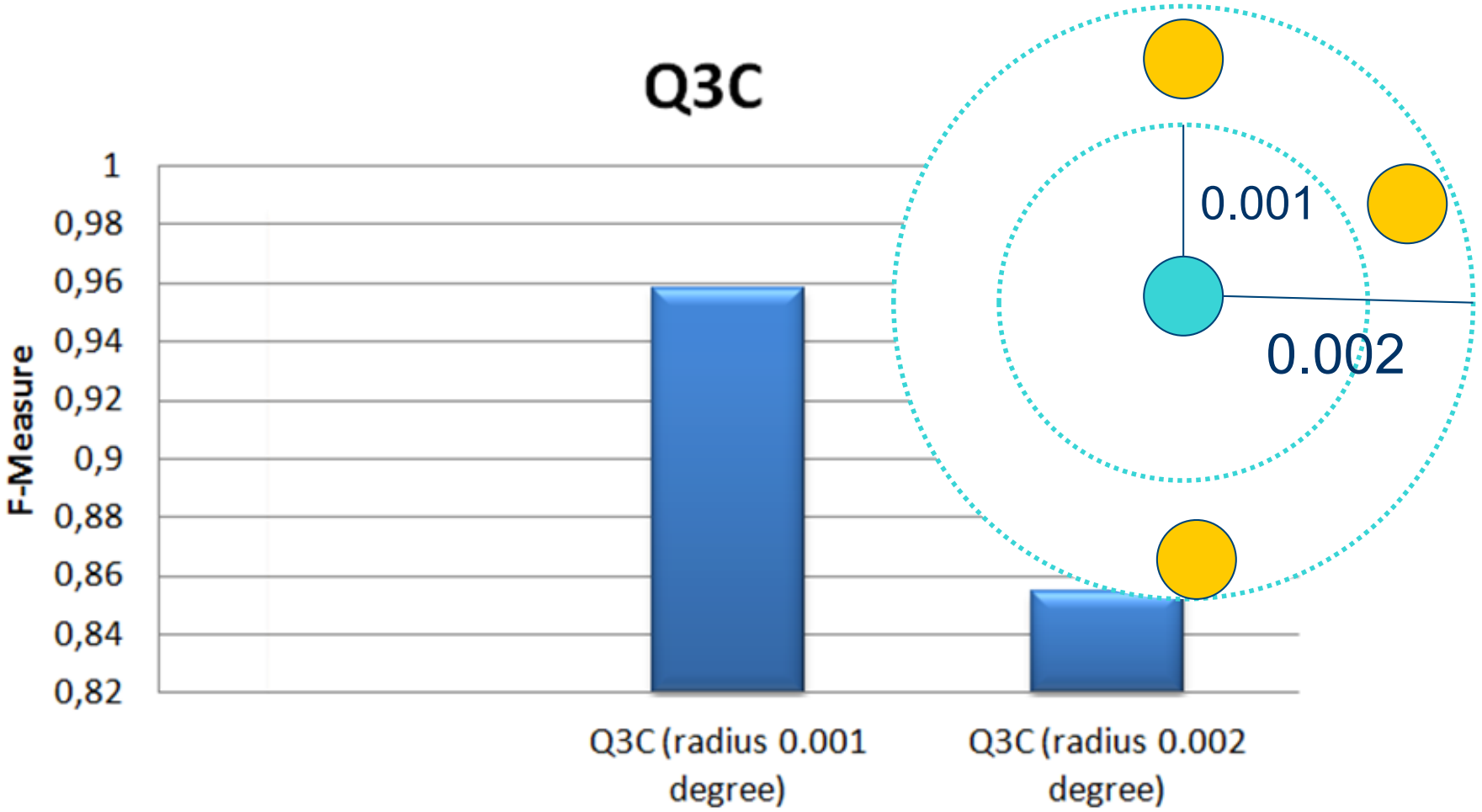
Comparasion with Q3C Join



- **Q3C Join**
  - Binary Cross-matching
  - Output: a catalog containing the matched objects
- **Comparasion**
  - Output of the Q3C Join execution with the output of the clustering algorithm NACluster
  - Input: the same two catalogues
    - Part of 2Mass catalog (1 million objects and the synthetic catalog generated from it (1 million objects)

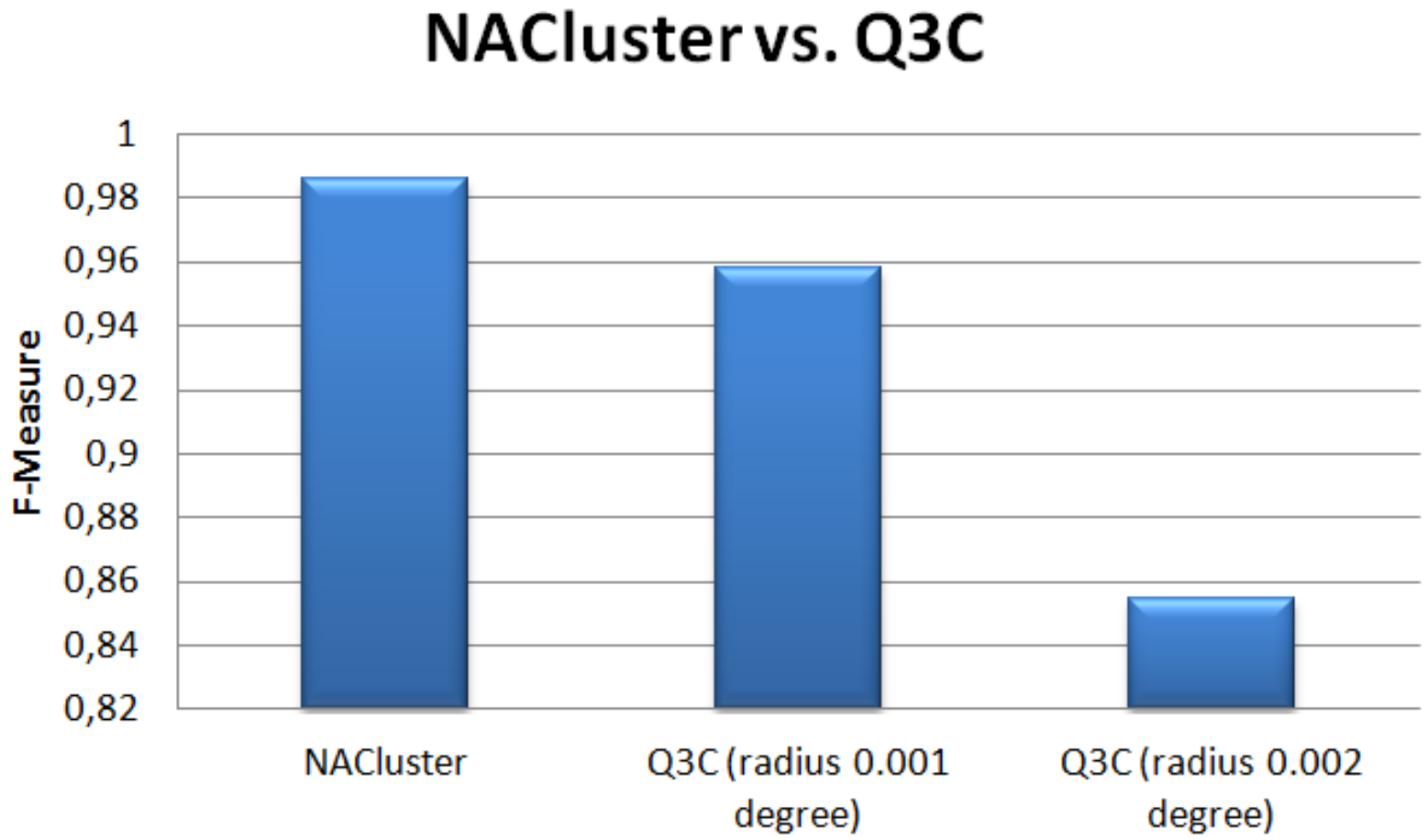
DEXL LAB
EXTREME DATA LAB

NACluster vs. Q3C Join



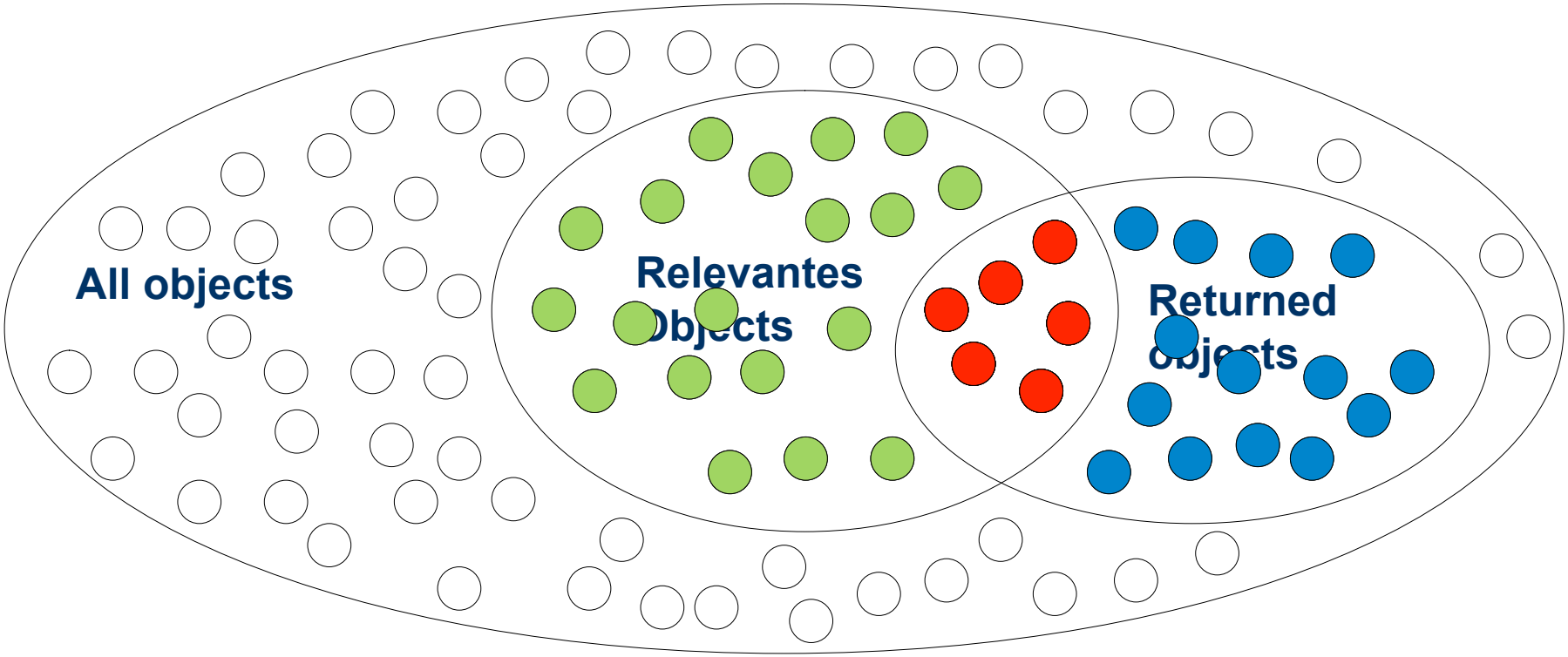
DEXL LAB
EXTREME DATA LAB

NACluster vs. Q3C Join



DEXL LAB
EXTREME DATA LAB

Precision, Recall and F-Measure



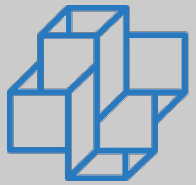
$$F = \frac{2 \times \text{Precisão} \times \text{Abrangência}}{\text{Precisão} + \text{Abrangência}}$$

DEXL LAB
EXTREME DATA LAB

Pseudocode

- When d(Oi;Ca) < ϵ, the object Oi is candidate to map to cluster Ca.
  - This mapping, however, can only be applied when that distance is the shortest distance and there not exists another object Oj in cluster Ca that has been mapped to the same catalog of Oi.

Conclusion



- These preliminary results indicate that the algorithm is effective in matching objects from different catalogues.
- Now we are developping a parallel strategy for NACluster algorithm using a spatial indexing strategy in order to reduce the complexity.

Pseudocode

- In case an object of Ca already exists in the cluster, two scenarios must be evaluated:
  - **(1)** if d(Oj ;Ca) > d(Oi;Ca) then we should remove the object Oj from the cluster Ca, insert Oi in this cluster, and search another cluster for Oj ;
  - (2) if d(Oj ;Ca) < d(Oi;Ca) then the algorithm performs a recursive search on the centroid candidate list for allocating Oi.
- In case, no cluster is found at distance epsilon then a new cluster Cb is created to the point Oi and it will be the centroid.