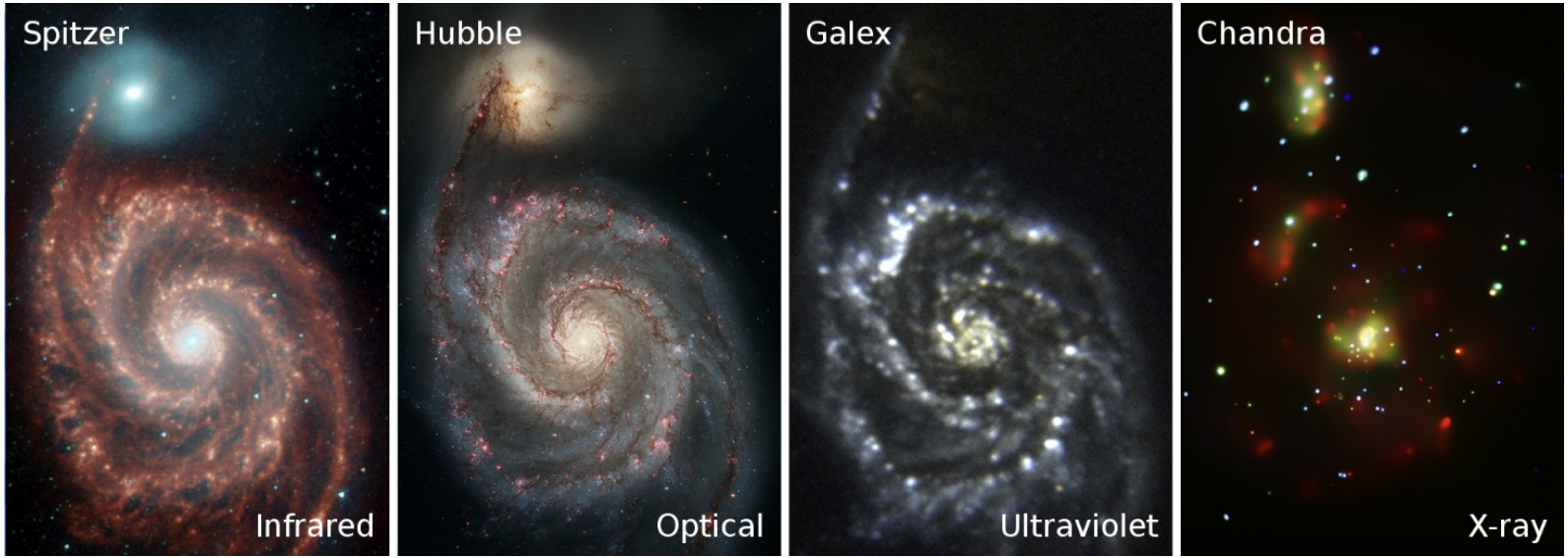
JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

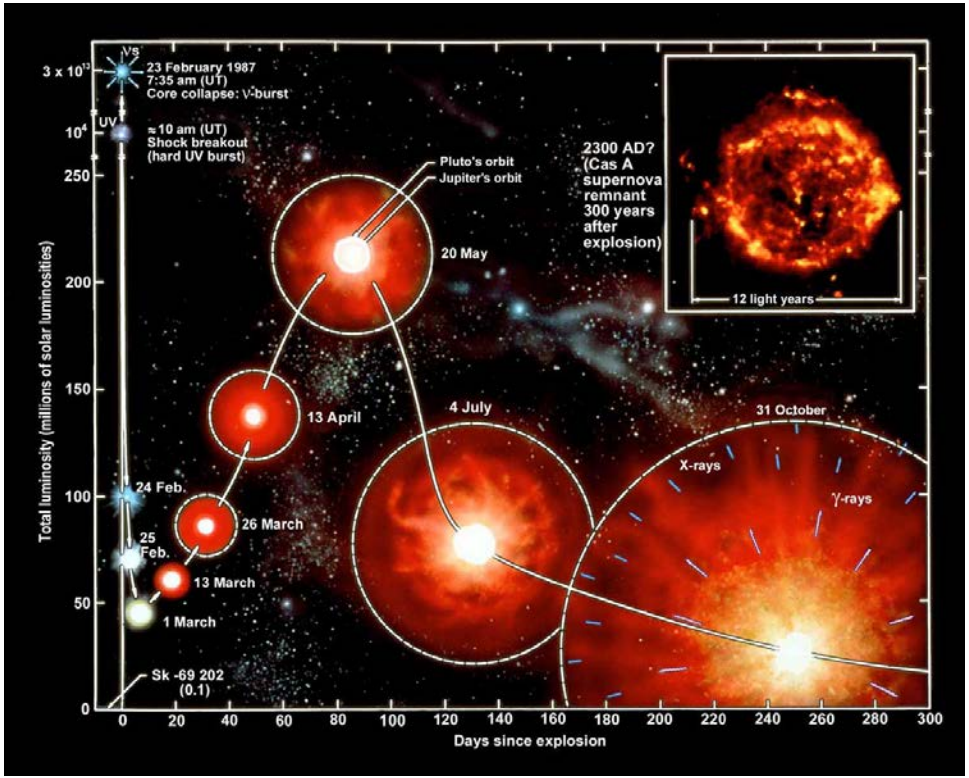
Towards Probabilistic Catalog Matching

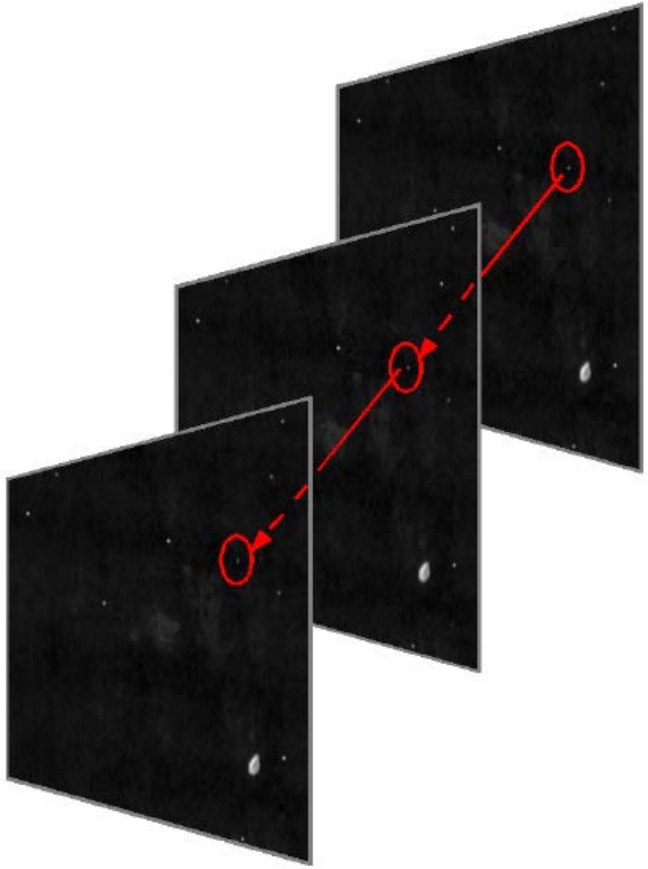
Tamás Budavári
Dept of Applied Mathematics & Statistics
Dept of Computer Science
Dept of Physics & Astronomy

Multicolor Universe



Eventful Universe





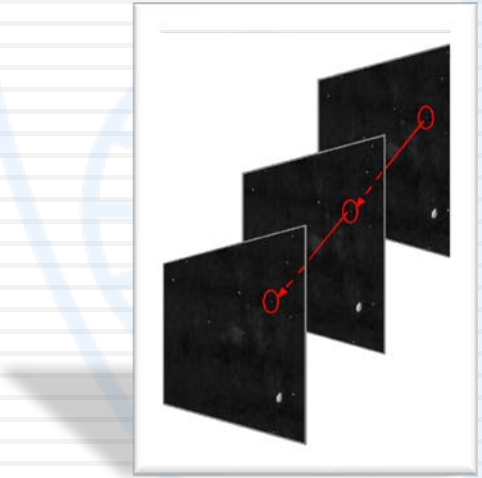
Fundamental Challenges

- Astronomy became statistical
  - Homogenous datasets from surveys
- Data fusion across instruments & epochs
  - For time-domain, multicolor & -messenger studies

Outline

- Goodness of an association
- Optimal catalogs of matches
- Open questions

Matching Sources



How?

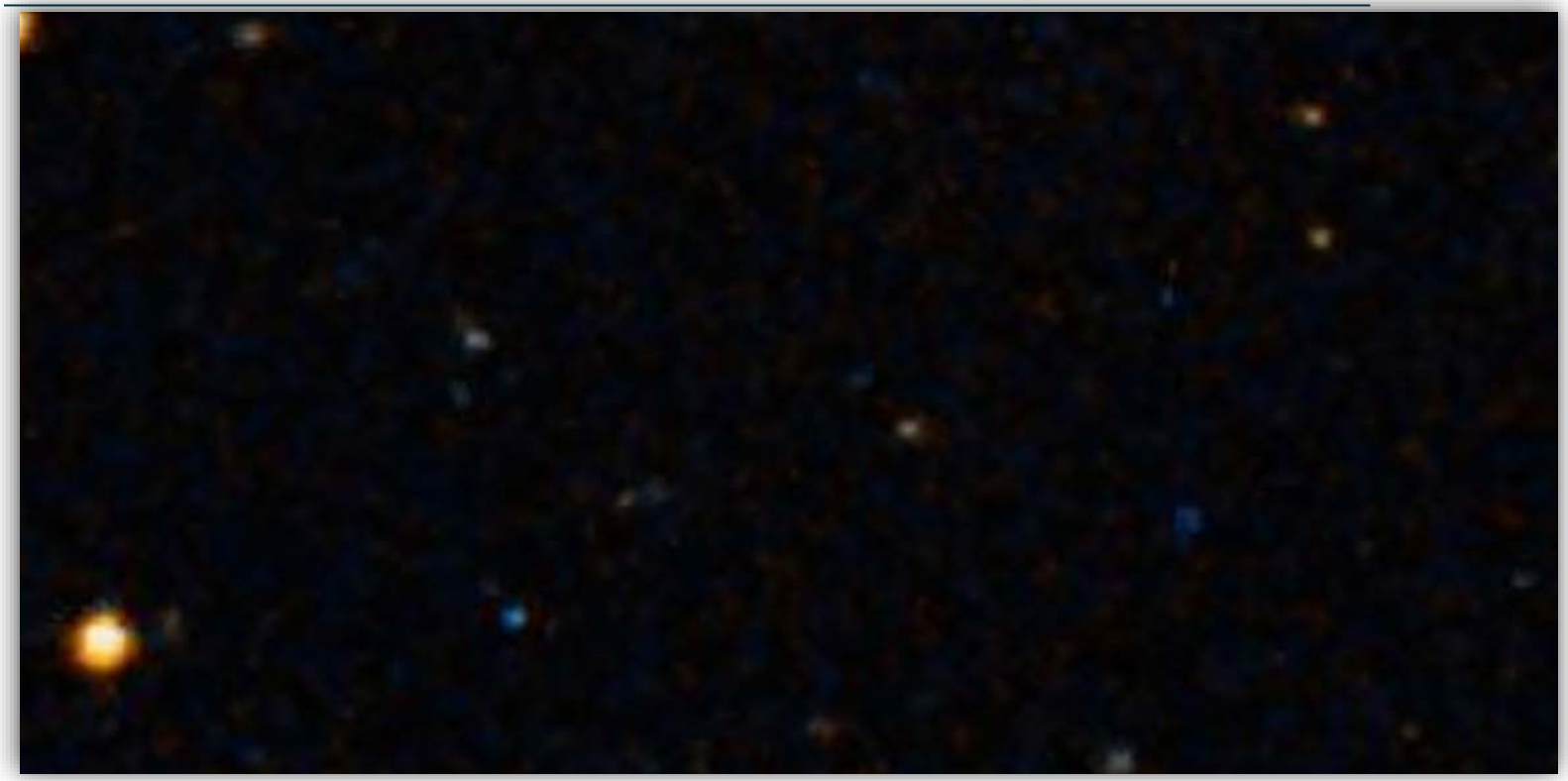
- Angular separation of nearby detections?

How?

- Angular separation of nearby detections?
  - Variable uncertainties
- Large contrast in surface density

How?

Digitized Sky Survey



How?

Hubble UDF

How?

- Angular separation of nearby detections?
  - Variable uncertainties
- Large contrast in surface density
  - Many-to-many candidates
  - Friends-of-friends chains
- Selection functions also differ!

Many methods

- Astro applications use various approaches
  - Differences inspired many asymmetric algorithms
    - Particular ordering for considering catalogs

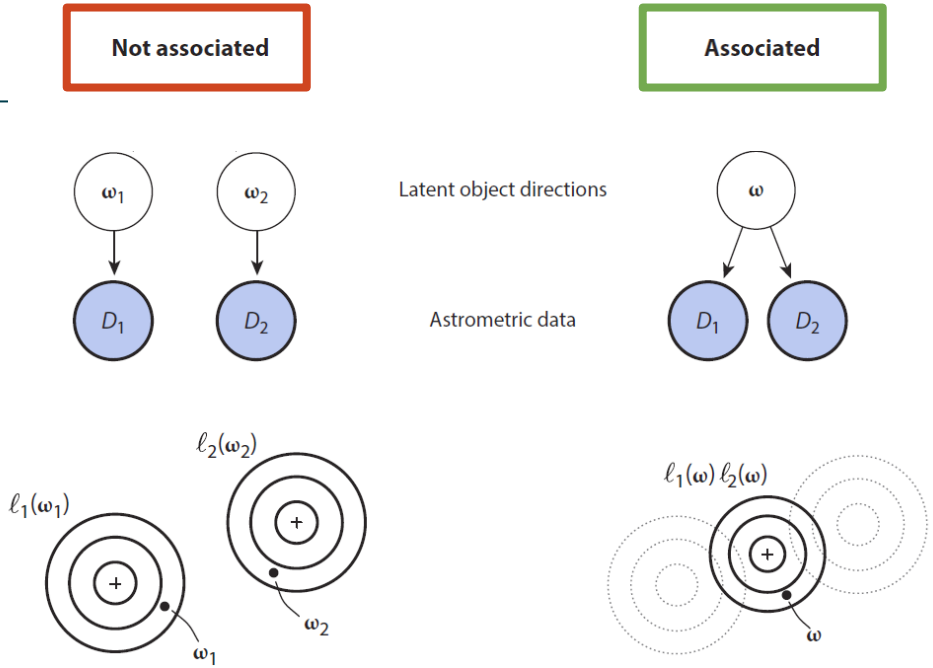
Definitions

- Source *(i,c)*
  - Entry *i* in catalog *c*
    - Detection with uncertain direction, etc.
- Object *(o)*
  - A celestial body we might observe
    - Hypothesized direction and properties

Bayesian

- Hypothesis testing: same or not
- Compare marginal likelihoods

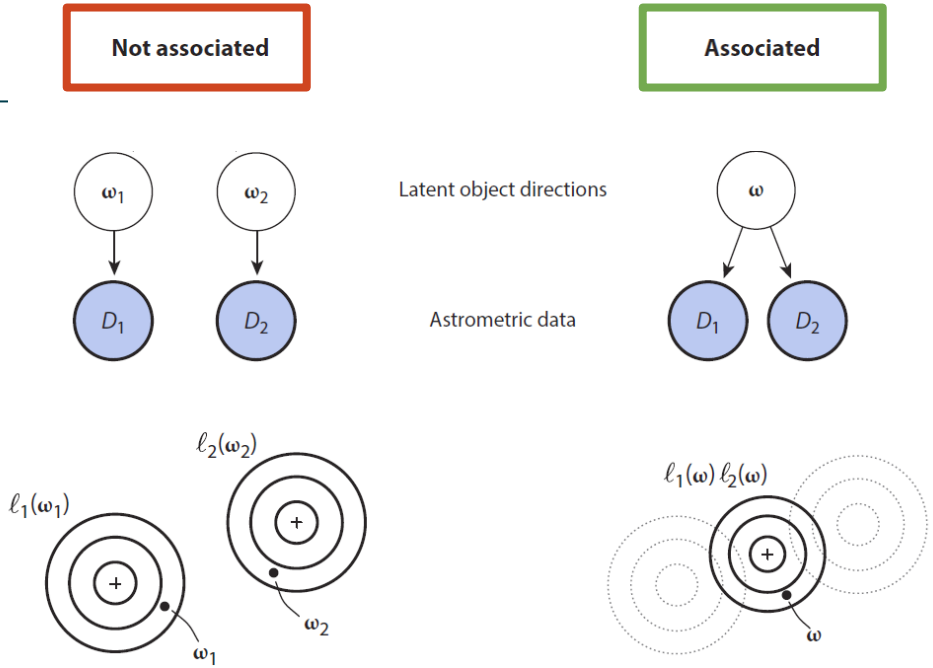
$$B_o = \frac{\mathcal{M}_o}{\mathcal{M}_o^{\mathrm{NA}}}$$



Bayesian

- Hypothesis testing: same or not
- Compare marginal likelihoods

$$B_o = \frac{\mathcal{M}_o}{\mathcal{M}_o^{\mathrm{NA}}}$$



$$\mathcal{M}_o^{\mathrm{NA}} = \prod_{(i,c)\in S_o} \int d\omega\ \rho_c(\omega)\ell_{ic}(\omega) \qquad \mathcal{M}_o = \int d\omega\ \rho_{C_o}(\omega) \prod_{(i,c)\in S_o} \ell_{ic}(\omega)$$

Normal Distribution

- Astrometric precision $\kappa_i = \frac{1}{\sigma_i^2}$
- Fisher distribution $f(\boldsymbol{x}; \boldsymbol{\omega}, \kappa) = \frac{\kappa}{4\pi \sinh \kappa} \exp\left(\kappa\, \boldsymbol{\omega}\boldsymbol{x}\right)$
  - Analytic results

$$B = \frac{\sinh \kappa}{\kappa} \prod_i \frac{\kappa_i}{\sinh \kappa_i} \quad \text{with} \quad \kappa = \left| \sum_i \kappa_i x_i \right|$$

Normal Distribution

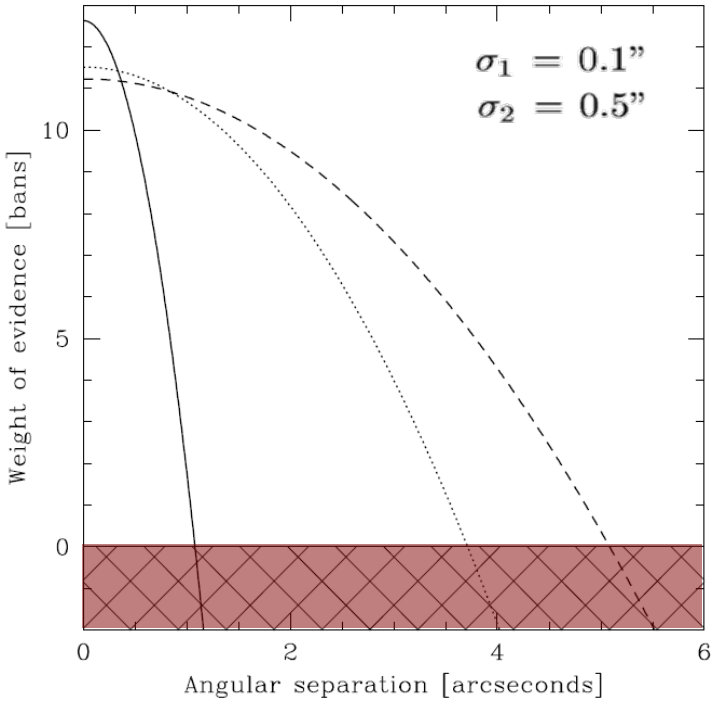
- *n*-way

$$B = 2^{n-1} \frac{\prod w_i}{\sum w_i} \exp\left\{ -\frac{\sum_{i<j} w_i w_j \psi_{ij}^2}{2 \sum w_i} \right\}$$

- *2*-way

$$B = \frac{2}{\sigma_1^2 + \sigma_2^2} \exp\left\{ -\frac{\psi^2}{2(\sigma_1^2 + \sigma_2^2)} \right\}$$

TB & Szalay (2008)



From Priors to Posteriors

- Posterior probability from prior & Bayes factor

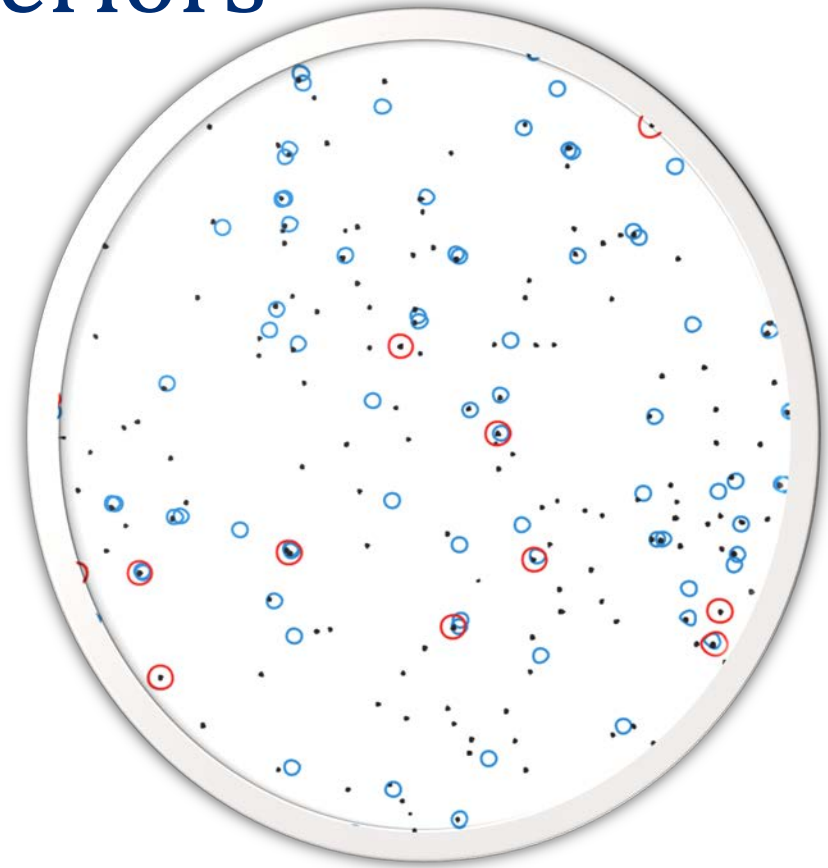
$$P(H|D) = \left[1 + \frac{1 - P(H)}{B\,P(H)}\right]^{-1}$$

- Prior probability of a match
  - Items in a bag: $1/N$
  - In general?

From Priors to Posteriors

- Different selections
  - **Nearby** / Distant
  - Red / Blue
- But only 1 number

$$P(H) = \frac{N_\star}{N_1 N_2}$$

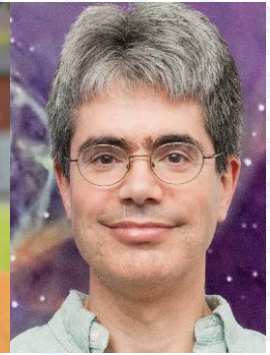


Successful Applications

- Include realistic galaxy clustering *(Mallinar, TB+ 2017)*
- Find moving stars w/unknown motion *(Kerekes+ 2010)*
- Events in time, supernovae *(TB 2011)*
- Radio lobes for EMU *(Fan, TB+ 2014)*
- X-ray for eROSITA *(Salvato+ 2017)*
- Transients to Hosts *(Aggarwal+ 2021)*

Match Catalogs! Not Sources…

With Amitabh Basu, Tom Loredo, et al.



Probabilistic Record Linkage in Astronomy: Directional Cross-Identification and Beyond

Tamás Budavári[1] and Thomas J. Loredo[2]

[1]Department of Applied Mathematics and Statistics, The Johns Hopkins University, Baltimore, Maryland 21218; email: budavari@jhu.edu

[2]Center for Radiophysics and Space Research, Cornell University, Ithaca, New York 14853; email: loredo@astro.cornell.edu

Annu. Rev. Stat. Appl. 2015. 2:113–39

The *Annual Review of Statistics and Its Application* is online at statistics.annualreviews.org

This article's doi: 10.1146/annurev-statistics-010814-020231

Keywords

partition models, directional statistics, astronomy, hierarchical Bayes, coincidence assessment

Abstract

Modern astronomy increasingly relies upon systematic surveys, whose dedicated telescopes continuously observe the sky across varied wavelength

Try Every Combination

- Imagine a simple 2-way case
  - Catalog-1
    - 2 sources
  - Catalog-2
    - 1 source

$$\mathcal{P}_0 = \left\{ \{(1,1)\},\ \{(2,1)\},\ \{(1,2)\} \right\}$$

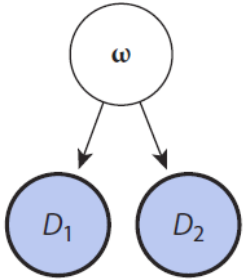
$$\mathcal{P}_1 = \left\{ \{(1,1)\},\ \{(2,1),(1,2)\} \right\}$$

$$\mathcal{P}_2 = \left\{ \{(1,1),(1,2)\},\ \{(2,1)\} \right\}$$

Likelihood of a Matched Catalog

- Marginal likelihood of a matched object *o*

$$\mathcal{M}_o = \int d\boldsymbol{\omega} \;\; \rho_{\mathcal{C}(o)}(\boldsymbol{\omega}) \prod_{(i,c)\in S_o} \ell_{ic}(\boldsymbol{\omega})$$



- Product over all objects

$$\mathcal{L}(\mathcal{P}) \equiv p(D|\mathcal{P}) = \prod_{o\in O_{\mathcal{P}}} \mathcal{M}_o$$

DIFFICULT COMBINATORIAL OPTIMIZATION!

Likelihood of a Matched Catalog

PROBABILISTIC CROSS-IDENTIFICATION IN CROWDED FIELDS AS AN ASSIGNMENT PROBLEM

Tamás Budavári[1,2,3] and Amitabh Basu[1]

ABSTRACT

One of the outstanding challenges of cross-identification is multiplicity: detections in crowded regions of the sky are often linked to more than one candidate associations of similar likelihoods. We map the resulting maximum likelihood partitioning to the fundamental assignment problem of discrete mathematics and efficiently solve the two-way catalog-level matching in the realm of combinatorial optimization using the so-called Hungarian algorithm. We introduce the method, demonstrate its performance in a mock universe where the true associations are known, and discuss the applicability of the new procedure to large surveys.

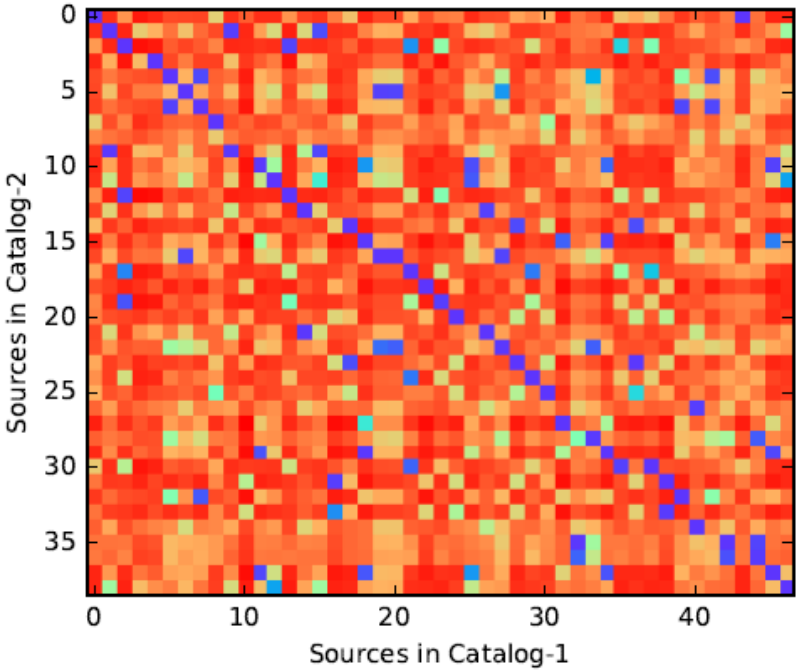
$$\mathcal{L}(\mathcal{P}) \equiv p(D|\mathcal{P}) = \prod_{o \in O_{\mathcal{P}}} \mathcal{M}_o$$

DIFFICULT COMBINATORIAL OPTIMIZATION!

Solvable for 2-way

- Assignment problem
  - E.g., workers to jobs
- Minimize overall cost
  - Rows – Columns

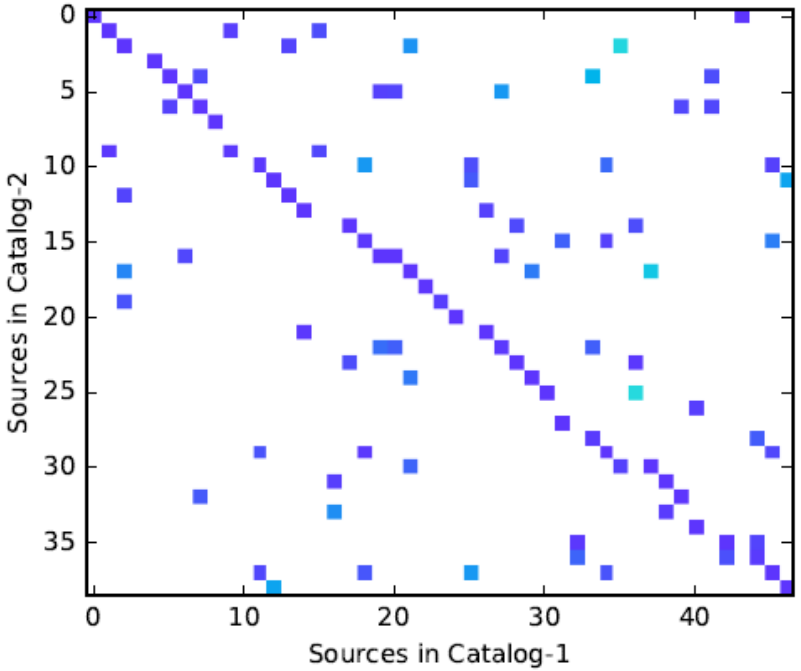
The Hungarian Algorithm



Solvable for 2-way

- Assignment problem
  - E.g., workers to jobs
- Minimize overall cost
  - Rows – Columns

The Hungarian Algorithm



Integer Linear Programming

- ILP to minimize –*log* likelihood of the catalog
  1. Binary variables to switch on/off candidates *(Shi+`18)*
  2. Or to assign detections to objects *(Nguyen+ `22)*

Integer Linear Programming

- ILP to minimize –*log* likelihood of the catalog
  1. Binary variables to switch on/off candidates *(Shi+`18)*
  2. Or to assign detections to objects *(Nguyen+ `22)*

Probabilistic Cross-identification of Multiple Catalogs in Crowded Fields

Xiaochen Shi[1,2], Tamás Budavári[1,3,4], and Amitabh Basu[1,3]

Published 2019 January 7 • © 2019. The American Astronomical Society. All rights reserved.

The Astrophysical Journal, Volume 870, Number 1

Citation Xiaochen Shi *et al* 2019 *ApJ* **870** 51

OPEN ACCESS

Globally Optimal and Scalable *N*-way Matching of Astronomy Catalogs

Tu Nguyen[1], Amitabh Basu[1,2], and Tamás Budavári[1,2,3]

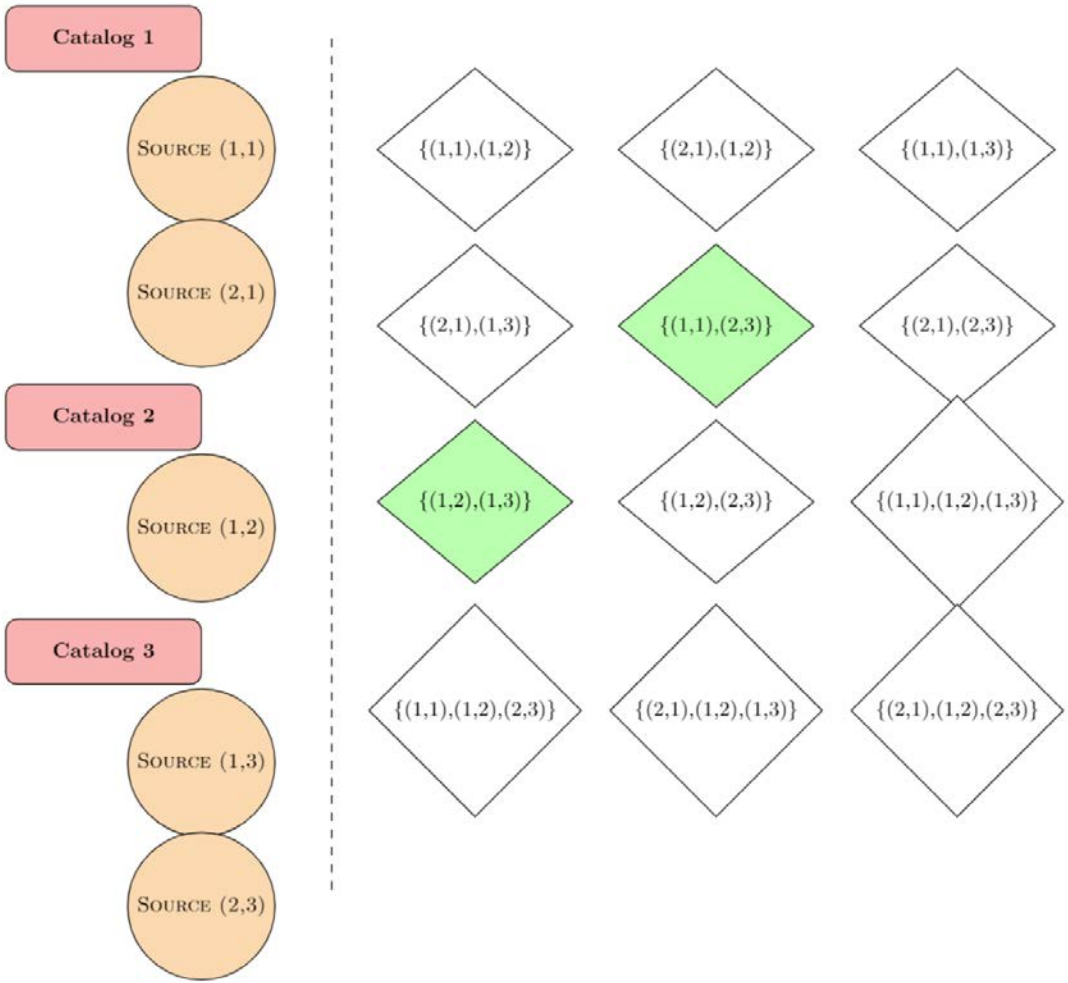
Published 2022 May 30 • © 2022. The Author(s). Published by the American Astronomical Society.

The Astronomical Journal, Volume 163, Number 6

Citation Tu Nguyen *et al* 2022 *AJ* **163** 296

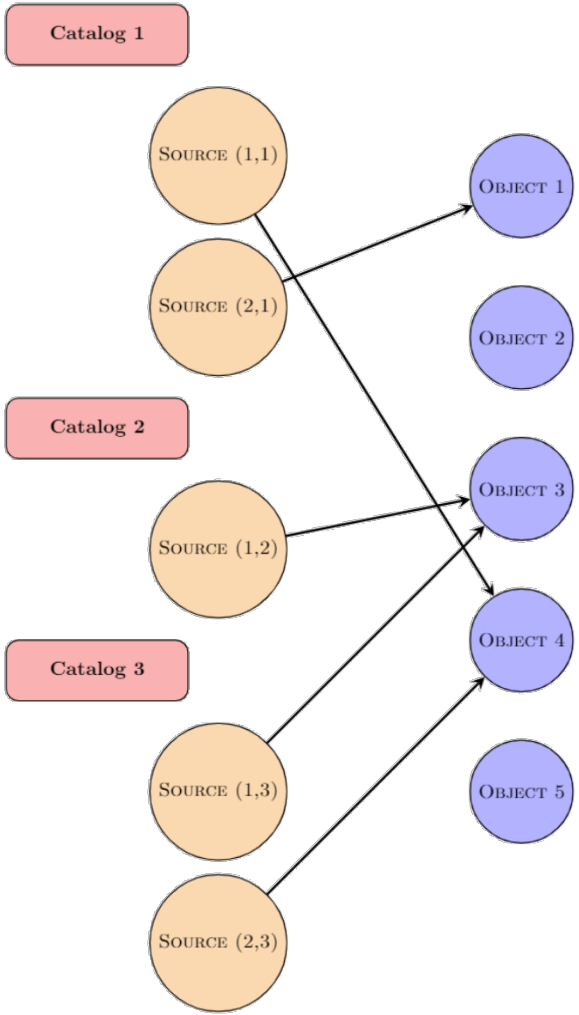
*Can*ILP

- First list all candidates
- ILP to find the optimal collection

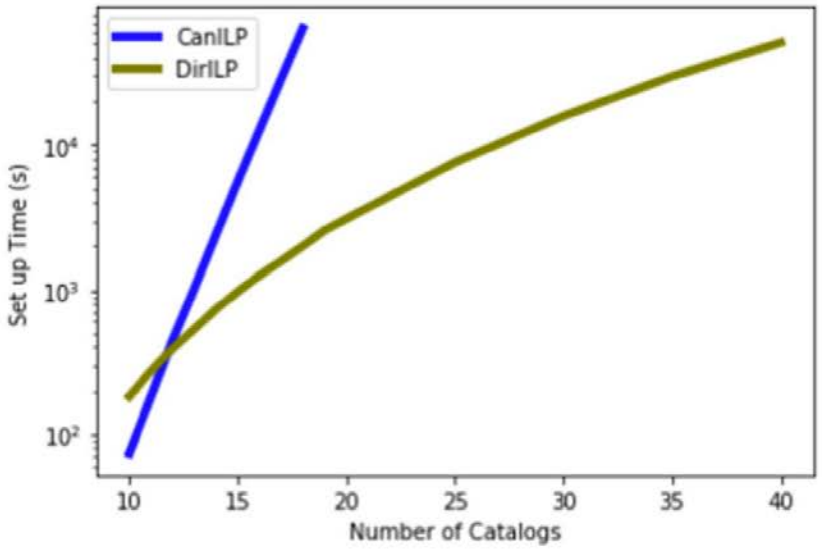


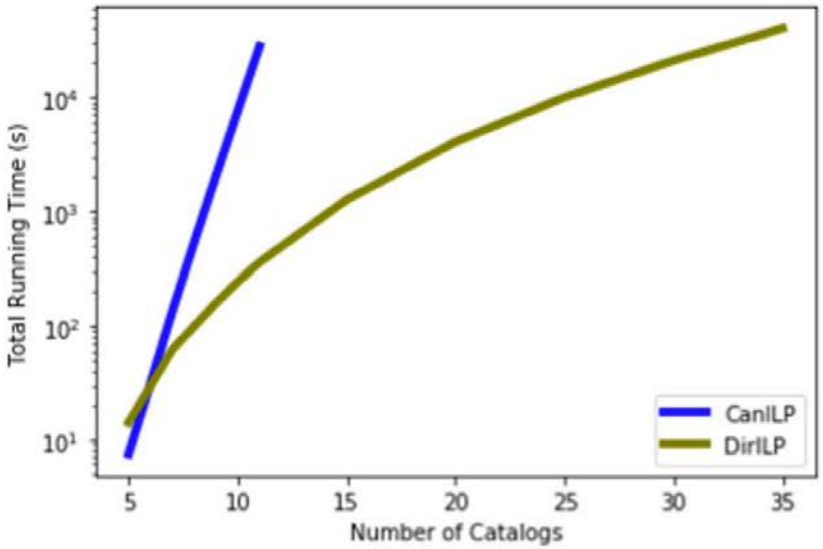
*Dir*ILP

- Direct assignment to hypothesized objects



Better Scaling



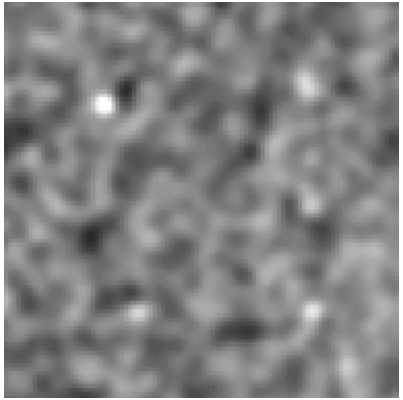


Summary

- Bayesian framework w/many applications
  - Computational tricks in practice
- Globally optimal solutions by ILP
  - *n*-way assignment problem
- Lots of open questions…

Some Open Questions

- Priors on match catalogs, e.g., # of objects?
- Data beyond directions, $B_{\text{flux,pos}} = B_{\text{flux}} \cdot B_{\text{pos}}$?
- Heuristics for more speed?
- Other "match" definitions?
  - Galaxy in cluster? Star in blend?
- Is a previous approach better?



JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING