Probabilistic Cross-identification of Multiple Catalogs in Crowded Fields

Xiao Chen Shi1,2 ©, Tamás Budavári1,3,4 ©, and Amitabh Basu1,3 ©

1 Department of Applied Mathematics and Statistics, The Johns Hopkins University, USA
2 Department of H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, USA
3 Department of Computer Science, The Johns Hopkins University, USA
4 Department of Physics and Astronomy, The Johns Hopkins University, USA

Received 2017 October 18; revised 2018 November 7; accepted 2018 November 9; published 2019 January 7

Abstract

Matching astronomical catalogs in crowded regions of the sky is challenging both statistically and computationally due to the many possible alternative associations. Budavári & Basu modeled the two-catalog situation as an assignment problem and used the famous Hungarian algorithm to solve it. Here we treat cross-identification of multiple catalogs by introducing a different approach based on integer linear programming. We first test this new method on problems with two catalogs and compare with the previous results. We then test the efficacy of the new approach on problems with three catalogs. The performance and scalability of this approach is discussed in the context of large surveys.

Key words: astrometry – catalogs – galaxies: statistics – methods: statistical – stars: statistics – surveys

1. Motivation

Dedicated telescopes systematically survey the night sky, often observing the same area. These exposures together enable multi-wavelength and time-domain studies. A central problem often observing the same area. These exposures together enable the development of statistical and computational methods. Based on Bayesian hypothesis testing, Budavári & Szalay (2008) developed a framework which proved to be applicable in a variety of scenarios, e.g., to account for proper motion (Kerekes et al. 2010), for radio morphology (Fan et al. 2015), or galaxy clustering (Mallinar et al. 2017); see also the review by Budavári & Loredo (2015) and references therein. New methods were also created to accelerate and automate the matching process on the largest catalogs available using spatial indices, e.g., the hierarchical triangular mesh (HTM; Kunszt et al. 2001) or HEALPix (Górski et al. 2005) within databases (e.g., the SkyQuery, Budavári et al. 2013) and even on graphics processing units (Lee & Budavári 2017). While these methods have opened the door to statistically sound data fusion techniques, they considered associations in isolation and simply assessed their quality without a global context.

In Budavári & Basu (2016), a non-trivial step beyond previous approaches was taken by introducing methods that maximized the likelihood of a global catalog matching, as opposed to a greedy choice of local likelihood maximization of isolated tuples. This was achieved by borrowing tools from combinatorial optimization which provide efficient algorithms for solving this much more difficult global optimization problem. Even so, the approach of Budavári & Basu (2016) can only be applied to matching two catalogs at a time, which is a serious limitation, as matching more catalogs is often needed in practice. In these situations, only heuristic methods have been available that find solutions based on greedy, locally good associations, and may not provide globally optimal matchings. This serious defect results in poor performance, especially when matching a large number of catalogs in crowded fields.

In this paper, we introduce a new class of algorithms that efficiently solves the problem of associating multiple catalogs while guaranteeing global optimality with respect to precise statistical objectives. Section 2 discusses the underlying partition models, the likelihood functions, marginal likelihoods, and Bayes factors, as well as our new optimization procedure. In Section 3, we study realistic simulations to test the accuracy and applicability of the new method. We also compare with the results of Budavári & Basu (2016) using the Hungarian algorithm (Munkres 1957). In Section 5, we demonstrate the improvements achieved by the new approach on three catalogs using a suite of simulations.

2. Catalog-level Matching

Probabilistic catalog-level cross-identification aims at finding the globally optimal catalog matches in which all associations are valid and no detection appears in multiple matches. Following Budavári & Loredo (2015), we discuss the formal model for the association problem, and introduce the statistical objective we shall use to find the best solution.

2.1. The Partition Model

We consider the union of all observations across the different catalogs and aim to partition this set of all observations into disjoint subsets. A subset with multiple elements should be interpreted as representing the hypothesis that all these elements in the subset are observations of the same object in the sky; such a subset is called an association. Thus, a subset containing a single element is to be thought of as the only observation among all the catalogs for that object; such a singleton association is called an orphan. Clearly, the partitioning cannot be arbitrary: any subset in the partition should contain at most one observation from each catalog (representing the fact that all these elements are observations of the same object in the sky). The fact that we have a partition into disjoint subsets encodes the fact that an observation should not stand for two different objects in the sky.
Let us consider a simple case. Suppose there are three catalogs, and they contain three, two, and four sources respectively. The \((i, c)\) pair denotes the \(i\)th source in catalog \(c\). Let \(D\) denote the set of all observations \((i, c)\), that is \(D = \{(i, c) : \forall i, c\}\). Then a matching between the three catalogs corresponds to a partition of \(D\) into disjoint subjects satisfying the constraint outlined in the previous paragraph. For example,

\[
\mathcal{P}_0 = \{\{(1, 1), (2, 2), (1, 3)\}, \{(3, 1), (4, 3)\}, \\
\{(1, 2), (3, 3)\}, \{(2, 1)\}, \{(2, 3)\}\}
\]

\[
\mathcal{P}_1 = \{\{(1, 1), (2, 3)\}, \{(3, 1), (3, 3)\}, \{(2, 1), (2, 2)\}, \\
\{(1, 2)\}, \{(1, 3)\}, \{(4, 3)\}\}
\]

\[
\mathcal{P}_2 = \{\{(1, 1), (2, 2), (3, 3)\}, \{(3, 1), (1, 2), (4, 3)\}, \\
\{(2, 1), (1, 3)\}, \{(4, 3)\}\}
\]

where \(\mathcal{P}_0\) gives five associations representing five objects (with two orphans), \(\mathcal{P}_1\) gives six associations for six objects (with three orphans); see Figure 1 and \(\mathcal{P}_2\) represents five separate objects (with three orphans). As mentioned earlier, the partition is valid if and only if any association in the partition has at most one of each catalog. Formally, if it contains a pair \([(i_1, c_1), (i_2, c_2)]\), then \(c_1 \neq c_2\). In this example, there are a total of seven valid partitions. We will use \(o \in \mathcal{O}_P\) to denote a collection of indices that index the subsets in a valid partition \(\mathcal{P}\), corresponding to objects in \(\mathcal{P}\). The catalog cross-identification problem is then equivalent to finding the optimal partition that has the maximum likelihood.

2.2. Probabilistic Formalism

The probabilistic formalism for cross-identification based on Bayesian hypothesis testing lends itself to hierarchical modeling and catalog-level generalizations; however, until now, no approach has been available to efficiently solve the problem of partitioning the sources into globally optimal associations of objects.

Given a valid partition \(\mathcal{P}\), one considers the conditional probability density of the data \(D = \{(i, c)\}\) given \(\mathcal{P}\), i.e., \(p(D|\mathcal{P})\). The hierarchical Bayesian framework has a vital feature that, given a certain partition, the data from distinct associations are conditionally independent. Therefore, the likelihood function at a certain partition \(\mathcal{P}\) can be factored as

\[
\mathcal{L}(\mathcal{P}) \equiv p(D|\mathcal{P}) = \prod_{o \in \mathcal{O}_P} \mathcal{M}_o
\]

where \(\mathcal{M}_o\) is the marginal likelihood of object \(o \in \mathcal{O}_P\) in the partition \(\mathcal{P}\) (Budavári & Loredo 2015), given by

\[
\mathcal{M}_o = \int d\omega \; \rho_{c(o)}(\omega) \prod_{i \in o} \ell_{ic}(\omega)
\]

where \(S_o\) denotes the subset \(\{(i, c)\}\) in the partition \(\mathcal{P}\) associated with object \(o\), \(C(o)\) represents a collection of catalogs containing sources associated with object \(o\), \(\ell_{ic}(\omega)\) is the likelihood function of source \(i\) in catalog \(c\) for direction \(\omega\), and \(\rho_{c(o)}(\omega)\) is the prior probability density for \(C(o)\) depending on the angular and radial selection functions of catalogs. Our goal is to find the optimal partition by maximizing the \(\mathcal{L}(\mathcal{P})\) catalog-level likelihood.

The marginal likelihood for non-association given all observed sources in \(S_o\) is

\[
\mathcal{M}_o^{NA} = \prod_{(i, c) \in S_o} \int d\omega \; \rho(\omega) \ell_{ic}(\omega)
\]

where \(\rho(\omega)\) is the prior probability density of direction \(\omega\) for an “orphan” in catalog \(c\). The Bayes factor is the ratio of the marginal likelihoods for the association and non-association hypotheses, quantifying how much the data support a specific association \(o\),

\[
B_o \equiv \frac{\mathcal{M}_o}{\mathcal{M}_o^{NA}}.
\]

Large \(B_o\) denotes that the sources in \(S_o\) are more likely from the same object \(o\) than different objects, while \(B_o\) less than one indicates the sources are more likely from separate objects.

Considering that the product of the marginal likelihoods \(\prod \mathcal{M}_o^{NA}\) over all \(o \in \mathcal{O}_P\) consists of the same terms corresponding to every observed source \((i, c)\),

\[
\prod_{o \in \mathcal{O}_P} \mathcal{M}_o^{NA} = \prod_{(i, c) \in D} \int d\omega \; \rho(\omega) \ell_{ic}(\omega)
\]

its value is actually independent of the \(\mathcal{P}\) partition. This means that maximizing \(\prod B_o\) is equivalent to maximizing the product of the marginals, \(\prod \mathcal{M}_o\).

In practice, we have to choose a member likelihood function, which is often assumed to be Gaussian. A more general choice is the Fisher (1953) distribution, which is a spherical analog to the normal distribution,

\[
f(\chi; \omega, \kappa) = \frac{\kappa}{4\pi \sinh \kappa} \exp(\kappa \omega \chi)
\]

where \(\omega\) is the unit vector of the mode, and \(\kappa\) is the concentration parameter. We can apply this spherical normal distribution, which is a spherical analog to the at-sky approximation corresponds to using Gaussian likelihoods with \(\kappa = 1/\sigma^2\), and the Bayes factor becomes

\[
B_o = 2^{2|S_o|-\sum k \kappa_{ic}} \exp \left( -\sum_{i,c} \sum_{j,k} \kappa_{ic} \kappa_{jk} \psi^2_{ic,jk} / 4 \sum_{i,c} \kappa_{ic} \right)
\]

where the \(\sum k\) and both \(\sum ic\) and \(\sum jk\) go over \(S_o\) set, e.g., \((i, c) \in S_o\), the \(\kappa_{ic}\) is the concentration parameter that corresponds to the astrometric uncertainty as in Equation (7), and \(\psi_{ic,jk}\) is the angle between the directions of sources \((i, c)\) and \((j, k)\); see Budavári & Szalay (2008).

In preparation for the optimization, we first take the logarithm of the above objective function \(\prod B_o\), which yields a minimization problem of the sum

\[
F(\mathcal{P}) = -\sum_{o \in \mathcal{O}_P} \ln B_o
\]
over all possible partitions $\mathcal{P}$ that satisfy the condition stated earlier that if $(i_1, c_1)$ and $(i_2, c_2)$ belong to the same subset in the partition, then $c_1 \neq c_2$.

2.3. Solution by Integer Linear Programming

We will solve the problem of minimizing (9) by formulating it as an integer linear program (ILP). An ILP is a mathematical optimization model that consists of several decision variables whose values are constrained to take only integer values, and the variables are subject to linear equality or inequality constraints. Subject to these integrality and linear constraints, the goal is to find values for the variables that minimize a given linear function of these decision variables.

To model (9) as an ILP, we introduce variables for each possible valid subset that can be formed from the given data set $D$ as described in Section 2.1. For every nonempty subset $T \subseteq D$ that contains at most one element from each catalog, we introduce a variable $x_T$ that will be constrained to be integer, and also by the linear inequalities $0 \leq x_T \leq 1$. This forces $x_T$ to be a binary variable that takes only values 0 or 1. The interpretation of this is as follows: if $x_T = 1$ in a solution, then the subset $T$ is taken as a subset in the partition, and if $x_T = 0$, then the subset $T$ does not appear in the partition. To ensure that we indeed have a partition, we must impose the linear equality constraint $\sum x_T = 1$ for every element $(i, c) \in D$ where the summations run over all $T$ that contain $(i, c)$. This forces exactly one subset to contain the element $(i, c)$, and thus the collection of subsets $T$ such that $x_T = 1$ will form a valid partition. Note that we have already modeled the fact that if $(i_1, c_1)$ and $(i_2, c_2)$ are in the same subset then $c_1 \neq c_2$, by introducing decision variables only for those subsets of $D$ that contain at most one element from each catalog.

To summarize, any valid partition $\mathcal{P}$ corresponds to an assignment of values to the variables $x_T$ that satisfy the following constraints:

\[
x_T \in \{0, 1\}, \quad 0 \leq x_T \leq 1, \\
\sum_{T: (i, c) \in T} x_T = 1 \text{ for every } (i, c) \in D.
\] (10)

Conversely, any $x$ satisfying all of these constraints provides a valid partition as the collection of subsets $\{T: x_T = 1\}$.

The next step is to model the objective function $F(\mathcal{P})$ using our decision variables $x_T$. This is a simple task achieved by computing the constants $w_T = -\ln B_T$ for every subset $T$ (we abuse notation slightly by using $B_T$ to denote $B_0$, where $o$ is the “potential” object represented by the subset $T$). Then the objective we want to minimize is simply $\sum w_T x_T$.

As a final simplification step, we observe that the number of variables can be significantly reduced for efficiency. The key point is that $\ln B_0 = 0$ if the object $o$ is an orphan, which follows from the definition of the Bayes factor. Thus, these coefficients do not contribute to the objective function. So one can simply remove the variables $x_T$ such that $|T| = 1$, i.e., $T$ is a singleton, and replace the partition constraint $\sum x_T = 1$ in Equation (10) by $\sum x_T \leq 1$, where again the summation runs over all $T$ that contains $(i, c)$. In the final solution, if an element $(i, c)$ is not part of any subset $T$ such that $x_T = 1$, then $(i, c)$ should be considered an orphan. Thus, our final integer linear programming function for minimizing Equation (9) over all valid partitions $\mathcal{P}$ is

\[
\min \sum_T w_T x_T \\
\text{subject to } x_T \in \{0, 1\} \text{ and } 0 \leq x_T \leq 1 \text{ for all } T, \\
\text{and } \sum x_T \leq 1 \text{ for every } (i, c) \in D.
\] (11)

In practice, the number of valid partitions can be large but the integer programming algorithms can be further improved with simple thresholding. We can reduce the number of variables by making the observation that the objective coefficient $w_T$ for some subsets $T$ is so large that the corresponding variable cannot possibly be set to 1 in any optimal solution (since we are minimizing). This typically happens for a subset $T$ that contains sources which are very far away, thus producing a large $w_T$ value. So, by setting a heuristic threshold to rule out subsets with large $w_T$, we eliminate a large fraction of the variables and accelerate the integer programming algorithm. This pruning is critical in practical implementations, and even a threshold of $\ln B_o = 0$ can do the trick, which is the value for orphans; thus, if $w_T \geq 0$ for some $T$, one can create an equally good or better solution by setting all the sources in $T$ as orphans; see also Budavári & Basu (2016) for more details.

Many algorithms and off-the-shelf software solutions exist for ILP problems, which are efficient for large-scale problems. In this paper, we employ the commercial Inc. Gurobi Optimization (2016) package to test the efficacy of our approach. Our implementation is entirely in Python and interfaces with Gurobi via the gurobipy module to create the problem definition and to run the solver.

3. Mock Objects and Catalogs

We apply the ILP cross-identification procedure to realistic simulations of point sources to study the performance. We generate mock objects in a small field of view with a given density. Each object is assigned a random direction $\omega$ and a “physical property” represented by number $u$ chosen uniformly at random in $[0, 1]$; see also Heinis et al. (2009). Figure 2 shows these mocks with different colors representing the intrinsic properties.

Next we generate catalogs with directional errors considering astronomical uncertainties. For the illustrations, we take Gaussian errors with a nominal $0.1^\circ$ standard deviation. Because different instruments have different sensitivities, each catalog observes only a subset of objects. This is modeled by assigning a preset interval to each catalog, and only those objects with $u$ values in this interval appear in this catalog. The associations occur when catalogs have overlapping intervals,
a common range of selected objects. For example, the left panel of Figure 3 illustrates generated source catalogs with different selection functions. The directions of the detections are plotting in red, green, and blue, which scatter around the true direction shown by the “+” sign; see the detailed analysis below. First, however, we apply the new ILP algorithm to a suite of two-way matching problems and test it against the Hungarian algorithm proposed in Budavári & Basu (2016). As expected, the outcomes are exactly the same, since the two algorithms are optimizing the same objective. The statistical properties of these associations are discussed in detail by Budavári & Basu (2016).

3.1. Matching Three Catalogs

Matching three catalogs is much more challenging computationally and procedures such as the Hungarian algorithm are not available. Our new approach, however, works just the same because no assumption on the number of catalogs is needed for the model set up in (11).

We implement our approach in Python within a Jupyter Notebook. As mentioned above, we use the Inc. Gurobi Optimization (2016) package to solve the integer linear programs that arise from (11) for the three-catalog scenario.

In this application test, both objects density and selection function are accessible from the simulation process, making the prior probability of each two-way and three-way matching a known parameter; see the details in Equation (25) of Budavári & Szalay (2008). Thus, by choosing a conservative posterior probability threshold, we can easily compute the thresholds for the Bayes factor, producing a more accurate estimation. However, the situation is much harder for real data where neither priors nor posteriors are obtained. Budavári & Loredo (2015) discuss estimating priors and posteriors in this complicated scenario, which will be a direction for future study.

Figure 3 illustrates our results. The left panel shows the objects with “+” signs indicating the true directions on the sky, as well as the noisy (simulated) measurements in the three catalogs shown in red, green, and blue. The solid dots represent objects seen in multiple exposures. In the right panel, the “×” signs denote the directions of the estimated objects. The small solid black dots illustrate the positions of objects estimated as associations involving three sources, while the small solid dots in magenta correspond to two-way associations between every two catalogs. The rest are orphan associations.

3.2. Matching Multiple Catalogs

So far we have demonstrated the new method’s statistical accuracy, a series of tests are performed with our ILP algorithm and, as a comparison, a simpler greedy technique. This naive approach considers all candidate associations with sources closer than a predefined angular separation threshold, and selects the best one based on the average (or maximum) pairwise distances. The match is added to the output catalog and its sources are removed from the catalogs. One can repeat the above greedy step until no candidate associations are left. To make the naive method most competitive, we tested multiple thresholds from $4\sigma$ to $7\sigma$, where $\sigma$ is the uncertainty of the source direction, and chose the best case performing closest to the ground truth.

For each mock universe, objects are randomly generated in a $3' \times 3'$ field of view with density of 400 objects per square arcminute. We took $\sigma = 0.04''$ as the astronomical uncertainty to simulate three catalogs, which were matched by both the ILP and naive methods repeatedly. These numbers are somewhat arbitrary but motivated by the expected performance of upcoming surveys, such as the Large Synoptic Survey Telescope (LSST). In our tests the best naive threshold was $6\sigma$, which is what we adopted for the greedy procedure. Next we summarize the findings in three relevant scenarios that differ effectively by the selection functions of the different catalogs.

4. Improved Accuracy

To quantify the new method’s statistical accuracy, a series of tests are performed with our ILP algorithm and, as a comparison, a simpler greedy technique. This naive approach considers all candidate associations with sources closer than a predefined angular separation threshold, and selects the best one based on the average (or maximum) pairwise distances. The match is added to the output catalog and its sources are removed from the catalogs. One can repeat the above greedy step until no candidate associations are left. To make the naive method most competitive, we tested multiple thresholds from $4\sigma$ to $7\sigma$, where $\sigma$ is the uncertainty of the source direction, and chose the best case performing closest to the ground truth.

For each mock universe, objects are randomly generated in a $3' \times 3'$ field of view with density of 400 objects per square arcminute. We took $\sigma = 0.04''$ as the astronomical uncertainty to simulate three catalogs, which were matched by both the ILP and naive methods repeatedly. These numbers are somewhat arbitrary but motivated by the expected performance of upcoming surveys, such as the Large Synoptic Survey Telescope (LSST). In our tests the best naive threshold was $6\sigma$, which is what we adopted for the greedy procedure. Next we summarize the findings in three relevant scenarios that differ effectively by the selection functions of the different catalogs.

4.1. Identical Selection Functions

We first consider the idealized scenario of identical observations (and selection intervals), which yields the same sets of objects in the three generated catalogs. If we generated 3600 mock objects, there should be 3600 three-way associations formed from the three catalogs. Figure 4 presents the distribution of missing matches from both methods. The gray error bars represent the measurements of the naive method and their square root, which is close to a Poisson distribution with
sample mean $\lambda = 11.82$ shown by the solid gray bars. We see that this method almost never provides perfect results and it typically makes 11–12 mistakes in this small field of view. The results from the ILP method are shown in blue, and are clearly better. While the naive method yields more than 10 missing true matches 58% of the time, our new approach improve this to just 8%. Moreover, the sample mean of missing matches also improves to 5.79, which is a factor of 2 lower than the greedy procedure. The plot also shows that the ILP algorithm provides perfect matches more than 5% of the time.

Notable in the plot is that, for the ILP results, mistakes tend to occur in pairs and odd number errors appear in only 2% of iterations. This is actually caused by sparsity and uncertainty when generating the universe and catalogs. An odd number of errors can only occur when we associate observations corresponding to at least three or more nearby objects. When the universe is sparse, it is more likely to have a pair of close objects than three or more objects close to each other, thus making an odd number of mismatches much more unlikely than an even number.

We note that the above test as well as the following two assume perfect deblending, the ideal (yet impossible) scenario that corresponds to the worst case of source crowding. If the sources are blended on the images to the limit that the source extraction procedures cannot distinguish them, then both the greedy and the ILP methods improve in performance. This is studied in Section 5.2 in detail.

### 4.2. Variable Exposure Depths

Survey telescopes typically obtain observations with the same exposure times but the seeing (point-spread function (PSF)) changes from time to time, which yields a variable depth of the extracted source lists. We model this scenario by setting upper bounds of the selection intervals to 0.9, 0.95, and 1, while keeping the lower bounds at 0. The Venn diagrams in Figure 5 illustrate the different subsets due to the selection functions.

Figure 5 demonstrates the significant improvement of the ILP algorithm. In the left panel, the green error bars illustrate that, out of 180 true two-way matches, the new algorithm achieves perfect two-way matches 65% of the time, while the naive approach, shown by the cyan error bars, gets perfect matches less than 10% of the time. The sample mean also improves from 2.6 to 0.4, resulting in a drastic change in their approximated Poisson distributions, presented by green and cyan solid bars respectively. The right panel also shows better performance of the ILP approach in terms of the reported three-way matches, whose sample mean is 5.6 comparing with 10.5 from the naive method. Moreover, out of 3240 true three-way matches, our procedure has less than 10 missing true matches.
over 90% of the time. In comparison, the naive method achieves this accuracy only 50% of the time.

4.3. The General Case of Different Populations

Finally, we test the method’s accuracy in a more complex scenario that simulates completely different instruments and their observations with the different selection functions. We model this by different overlapping intervals producing 1080 two-way matches and 2050 three-way matches. Figure 6 presents these results along with the schematics of the selection functions using Venn diagrams. The left panel shows that, out of 1080 true two-way matches, the ILP approach produces perfect matches around 10% of the time. The plot also shows the approximating Poisson distribution with sample mean $\lambda = 2.7$. In contrast, the nearest-neighbor method includes no perfect matches and fits a flatter Poisson distribution with $\lambda = 16$. The right panel for three-way matches demonstrates that, when applying the new procedure, out of 2520 true three-way matches, around 4.5% of outputs are perfect matches and only 3.5% of results have more than 10 missing matches. The naive approach reports no perfect matches and produces more than 10 missing matches 30% of the time.

5. Discussion

The algorithm presented here is the first attempt at solving the general catalog-level association problem for multiple catalogs with provable guarantees. An objective function based on the marginal likelihood is constructed, which is then optimized to global optimality by modeling it as an ILP. In the previous sections we see a significant improvement due to the global optimization approach over the naive greedy method. Next we will discuss the feasibility of the approach by studying its performance and computational speed. We also look at the limit of imperfect deblending.

5.1. Runtime Analysis

For practical use, efficiency is considered as key to applicability. Budavári & Basu (2016) show that, for two-way catalog matching, we can simplify the problem by eliminating variables corresponding to highly unlikely associations, which in turn significantly reduces the runtime. The same holds for our new algorithm. In a direct comparison of wall-clock times, we find that our ILP approach method is over a factor of 50 faster than the Hungarian algorithm, which is most encouraging for large-scale applications. Of course, these differences could very well depend on the implementation—a more efficient implementation of the Hungarian algorithm should be more comparable to the ILP solver. Moreover, with an increasing number of objects, we expect the Hungarian algorithm to overtake the ILP approach in the two-way case.

For the three-catalog case discussed in this paper, the new algorithm has even more impressive performance. As seen in the left panel of Figure 7, the running time curve of the whole procedure, including pruning and optimization, scales...
quadtratically with the number of observations. This is primarily due to the inefficiency of our pruning procedure, which was not the focus of our study, as many approaches exist to speed this up, e.g., Kunszt et al. (2001), Górski et al. (2005), Gray et al. (2007), and Lee & Budavári (2017). In the right panel, we plot the runtime for the optimization only, which shows a clear linear scaling with the number of observations.

5.2. Imperfect Deblending

The previous tests assumed perfect deblending, which is both the ideal scenario for source detection and the worst case for crowding in the resulting catalogs. Here we look at how the results change in the case of blended sources as an illustration of the expected changes.

We adopt an approximate heuristic formula for our simulations based on the characteristics of extracted sources in the Sloan Digital Sky Survey using a nominal astronomical uncertainty of $\sigma \approx 0.1''$ and a PSF of FWHM $\approx 1.4''$. Keeping to this ratio, we scale to our previously assumed $\sigma = 0.04''$, which yields a PSF of FWHM $\approx 0.5''$. This corresponds to about 2.5 pixels in the LSST, our motivating example mentioned earlier. If we assume that sources closer than that separation on the sky cannot be deblended by the photometric software stack, we can simulate this effect by modifying the mock universe to include only one of the two colliding objects. Accordingly, the source catalogs now again have at most one source per astrophysical object (e.g., a star). This eliminates the ambiguity of the original blending issue when multiple stars could have a single blended source or photometric measurement. Of course, this is overly optimistic but our goal here is to study that limit, in addition to the previous (pessimistic) situation for cross-matching, the case of perfect deblending. The true expected performance is clearly between these two idealized scenarios.

To draw statistically significant conclusions, we created simulations in a larger area of 36 sq. arcmin, with the same object density as in Section 4, and re-ran the tests using the selection functions described in Section 4.1. Figure 8 illustrates the difference between a greedy algorithm and our optimal solutions in two extreme scenarios: the case of perfect deblending and the case of blended sources. We elaborate below.

Perfect deblending results in sources very close to each other on the sky, and hence crowding is the most challenging for correctly identifying the associations. The left panel of Figure 8 shows the performance in this scenario for the two approaches: the globally optimal solver producing more than two times fewer incorrect matches. The measurements are shown with blue and gray error bars for the ILP and greedy algorithms, respectively. In addition, we plot for reference the best-fit Poisson distribution. In particular, the greedy approach produces 48 wrong associations out of 14,400 objects on average, but it produces fewer than 40 bad matches only 15% of the time. In comparison, the ILP procedure achieves this same accuracy over 98% of the time and has a sample mean of 23 for missing true associations.

We model the situation with blending as follows: objects behind each other that are separated by an angle smaller than the PSF FWHM are not separated based on astrometry. We essentially eliminate the background object(s) and every catalog has at most one source corresponding to the blended objects. This obviously makes crowding less severe. In this limit both cross-identification approaches perform better as demonstrated in the right panel of Figure 8: the greedy algorithm produces 10 times better results than previously but only manages to produce perfect associations 0.6% of the time, which is in sharp contrast with our ILP method, which succeeds 99.5% of the time.

We expect the truth to lie between these two extremes but the fact that our novel optimal solver provides significantly better results in both cases suggests superior performance for real catalogs.

6. Concluding Remarks and Future Work

The approach based on the Bayesian probabilistic modeling and our proposed solution using ILP is potentially of broader applicability than what was discussed in the previous sections. First of all, the new algorithm can deal with $N$ catalogs simultaneously, as discussed in Section 2.3. Here we demonstrated the power of the technique on three catalogs.

Second, the ILP solver does not have to run on the entire Cartesian product of all catalogs. Instead it is meant to use a list of possible candidates found by existing high-performance tools (e.g., those based on the HTM, HEALPix, or the Zones Algorithm). The ILP solver is to only be invoked on sources in crowded or ambiguous regions where multiple competing matching scenarios are present.
Finally, our proposed method is applicable not only to static point sources (illustrated in this paper) but also to any other scenarios where one can model the morphology of objects and even their changes or proper motion (see the citations in Section 1). This is achieved by starting with the marginal likelihoods derived from the appropriate geometrical and physical models. Similarly, more advanced astrometric and population models will work, e.g., with elliptic errors or realistic clustering (Mallinar et al. 2017), as long as the marginal likelihoods \( \{M_n\} \) can be calculated. In other words, the new approach is a natural add-on package for all existing matching tools. These properties make it an ideal extension to existing pipelines and a must-have tool for modern surveys.

Our future work will include extensions to more complex settings, such as radio morphology (Fan et al. 2015), where even two-way catalog matches might comprise up to four sources (D. Fan et al. 2018, in preparation). On a conceptual level, further improvement is expected from going beyond the current maximum likelihood method. One can formulate simple priors for the different partition schemes by counting the number of associations with different numbers of constituents. Instead of treating these as priors, one can also consider these population probabilities in a hierarchical model, and infer them simultaneously.

T.B. acknowledges partial support from NSF Grant AST-1412566 and NASA via the awards NNG16PJ23C and STScI-49721 under NAS5-26555. A.B. acknowledges support from NSF Grant CMMI-1452820.

**ORCID iDs**

Xiaochen Shi @ https://orcid.org/0000-0002-6654-4807
Tamás Budavári @ https://orcid.org/0000-0002-7034-4621
Amitabh Basu @ https://orcid.org/0000-0002-1070-2626

**References**

Budavári, T., & Basu, A. 2016, AJ, 152, 86
Budavári, T., Dobos, L., & Szalay, A. S. 2013, CSE, 15, 12
Budavári, T., & Loredo, T. J. 2015, AnRSA, 2, 113
Budavári, T., & Szalay, A. S. 2008, ApJ, 679, 301
Fan, D., Budavári, T., Norris, R. P., & Hopkins, A. M. 2015, MNRAS, 451, 1299
Fisher, R. 1953, RSPSA, 217, 295
Górski, K. M., Hivon, E., Banday, A. J., et al. 2005, ApJ, 622, 759
Gray, J., Nieto-Santisteban, M. A., & Szalay, A. S. 2007, arXiv:cs/0701171
Heinis, S., Budavári, T., & Szalay, A. S. 2009, ApJ, 705, 739
Inc. Gurobi Optimization 2016, Gurobi optimizer reference manual, URL http://www.gurobi.com
Kerekes, G., Budavári, T., Csabai, I., Connolly, A. J., & Szalay, A. S. 2010, ApJ, 719, 59
Kunszt, P. Z., Szalay, A. S., & Thakar, A. R. 2001, in Mining the Sky, ed. A. J. Banday, S. Zaroubi, & M. Bartelmann (Berlin: Springer-Verlag), 631
Lee, M. A., & Budavári, T. 2017, A&C, 20, 155
Mallinar, N., Budavári, T., & Lemson, G. 2017, A&C, 20, 83
Munkres, J. 1957, SJAM, 5, 32

---

**Figure 8.** Performance of the greedy (gray points with error bars) and the optimal (blue) algorithms in the limiting cases with surface density 400/sq. arcmin simulated in a field of view of 36 sq. arcmin; the solid bars show the best-fit Poisson distributions for reference. The left panel illustrates the (impossible) scenario where all sources are perfectly deblended, such that the catalogs are completely crowded, which makes the association very hard. The right panel shows the scenario where we accept the fact that some blended sources will remain blended below a separation threshold, and hence we accept those as good matches; see the text. The true performance is expected to be between these two extremes.