Online Appendix of "Cross-identification of stellar catalogs with multiple stars: Complexity and Resolution"

Example of a 2-Matching Problem

Consider an instance of the 2-Matching Problem where $A = \{a_1, a_2, a_3, a_4\}$ and $B = \{b_1, b_2, \ldots, b_6\}$. Here, a_1, a_4 are single stars and a_2, a_3 are double. Suppose that the first phase yields the following sets:

$P_{a_1} = \{\{b_1\}, \{b_2\}, \{b_3\}\}$,
$P_{a_2} = \{\{b_2, b_3\}, \{b_4, b_5\}\}$,
$P_{a_3} = \{\{b_2, b_3\}, \{b_5, b_6\}\}$,
$P_{a_4} = \{\{b_6\}, \emptyset\}$.

A scheme that includes probabilities is displayed in Figure 3(a). Here, the optimal assignment is $f^*(a_1) = \{b_1\}$, $f^*(a_2) = \{b_4, b_5\}$, $f^*(a_3) = \{b_2, b_3\}$, $f^*(a_4) = \emptyset$ with probability $\overline{p}(f^*) = 0.1008$.

The reduction to the MWSSP gives $M = 8.3269$, weights

$w_{1\{1\}} = 7.1229$, $w_{1\{2\}} = 7.6338$, $w_{1\{3\}} = 6.7175$,
$w_{2\{2,3\}} = 7.1229$, $w_{2\{4,5\}} = 7.9702$,
$w_{3\{2,3\}} = 8.1038$, $w_{3\{5,6\}} = 6.7175$,
$w_{4\{6\}} = 7.4106$ and $w_{4\emptyset} = 7.8161$

(letters "a" and "b" are omitted for the sake of readability), and the graph is shown in Figure 3(b).

Example of the reduction of Lemma 3.1 and Theorem 3.2

Consider the instance of DM given in Figure 4(a) where $n = 2$ and $|\mathcal{A}_1| = |\mathcal{A}_2| = 3$. The corresponding instance of 2-MDPAW is shown in Figure 4(b) where $t = 2$.

Also, for $\beta = 0.4$ and the given instance of 2-MDPAW, the corresponding instance of 3-MDP is shown in Figure 5 where $t' = 3.6$. Vertices $\tilde{b}_{j3}$ for all $j \in \bigcup_{a \in A} P_a$ are displayed as unlabeled circles filled with white color.

Algorithm

Here, an exact algorithm for the the $\mathscr{K}$-Matching Problem is proposed and the resolution of a cross-identification between two catalogs based on real data is presented.

The algorithm is given below.

(i) For each $a \in A$ such that $\emptyset \in P_a$, do the following. If there is an element $j \in P_a$ such that $w_{aj} > w_{a\emptyset}$, remove j from P_a (if f^* is an optimal

assignment then $f^*(a) \neq j$ since $\emptyset$ is a better choice than j).

(ii) Generate the graph G as stated in Section 2.

(iii) Find the connected components of G.

(iv) For each component G' of G, solve the problem restricted to G'.

Let A' and B' be the stars involved in a component G' of G, i.e. $A' = \{a \in A : v_{aj} \in V(G')\}$ and $B' = \{b \in B : v_{aj} \in V(G'), b \in j\}$. In the last step of our algorithm, three cases can be presented:

- *Unique star.* If $A' = \{a\}$, then the solution is straightforward: $f^*(a) = argmin_{j \in P_a} w_{aj}$.
- *Only single stars.* If $|A'| \geq 2$ and $k_a = 1$ for all $a \in A$, then the problem restricted to G' can be solved via the Hungarian Algorithm in polynomial time. In that case, the instance of the MWMP is: a bipartite graph G_B such that $V(G_B) = A' \cup B' \cup \{\emptyset_a : a \in A' \text{ such that } \emptyset \in P_a\}$ and $E(G_B) = \{(a, b) : a \in A', \{b\} \in P_a\} \cup \{(a, \emptyset_a) : a \in A', \emptyset \in P_a\}$, weights $-w_{a\{b\}}$ for each edge (a, b) and weights $-w_{a\emptyset}$ for each edge $(a, \emptyset_a)$.
- *Multiple stars.* If $|A'| \geq 2$ and there is $a \in A'$ such that $k_a \geq 2$, then it can be solved with an exact algorithm for the MWSSP [6]. In the case that such algorithm is not available, solving the following integer linear programming formulation is a reasonably fast alternative:

$$\min \sum_{a \in A'} \sum_{j \in P_a} w_{aj} x_{aj}$$

subject to

$$\sum_{j \in P_a} x_{aj} = 1, \qquad \forall\, a \in A' \qquad (1)$$

$$\sum_{a \in A'} \sum_{j \in P_a : b \in j} x_{aj} \leq 1, \qquad \forall\, b \in B' \qquad (2)$$

$$x_{aj} \in \{0, 1\}, \qquad \forall\, a \in A',\ j \in P_a$$

Constraints (1) guarantee that each star of A' must be assigned to exactly one element j of P_a. Constraints (2) forbid that each star of B' be assigned to two or more stars of A'. For the sake of readability, the latter constraints are presented for all $b \in B'$ but one have to keep in mind that some of them can be removed if: (i) the constraint has just one variable in the left hand

[6] See, for instance, S. Rebennack, M. Oswald, D. O. Theis, H. Seitz, G. Reinelt and P. M. Pardalos, *A Branch and Cut solver for the maximum stable set problem*, J. Comb. Optim. **21** (2011), 434–457.

side, or (ii) it is repeated, i.e. if, for some $b \in B'$, there exists another $\tilde{b} \in B'$ such that b and $\tilde{b}$ occur exactly in the same tuples of $\bigcup_{a \in A'} P_a$.

An instance of the 2-Matching Problem is obtained once the first phase is completed. Table 1 reports some highlights about the optimization of that instance.

As we can see from the table, G is highly decomposable and just 111 integer linear problems needs to be solved. Moreover, these integer problems turned out to be very easy to solve since the solver did not branch (all of them were solved in the root node). The hardest one has 339 variables and 118 constraints, and took 0.0015 seconds of CPU time. The optimization was performed on a computer equipped with an Intel i7-7700 at 3.60 Ghz and GuRoBi 6.5.2 as the MIP solver. The overall process took 41.6 seconds of CPU time.

Description of the first phase

This section is devoted to present a summary on how to obtain a set of candidate stars for a given star of the former catalog and the probabilities involved in them. Recall that such computations heavily depends on structure and data availability of both catalogs as well as the underlying physical model used to establish the relationship between them. It is beyond the scope of this work to analyze such scenarios neither to give a formal treatment, so a simplified [7] but reasonable model is considered, which is enough for presenting our approach [8].

Consider catalogs A and B, and let A_2 be the set of stars from catalog A marked as "double". Our goal is to propose an instance of the 2-Matching Problem.

Let us first present some basic elements of Positional Astronomy. Usually, position is given in a well established reference frame where two spherical coordinates are used: *right ascension* denoted by α and *declination* denoted by δ, similar to longitude and latitude coordinates on Earth. In fact, a pair (α, δ) represents a point in the unit sphere. For a given two points p_1, p_2, denote its

[7] Stars of both catalogs should not be near the celestial poles in order to avoid certain distortions, and stars with high variability in its brightness should be avoided. This can be done by pre-identifying them and remove them from both catalogs.

[8] A more robust and general probabilistic model is discussed in T. Budavári and A. S. Szalay, *Probabilistic Cross-Identification of Astronomical Sources*, Astrophys. J. **679** (2008), 301–309.

angular distance by $\theta(p_1, p_2)$. A known property is that, if points p_1, p_2 have the same right ascension, $\theta(p_1, p_2)$ is given by the difference in its declinations. However, if p_1, p_2 have the same declination, $\theta(p_1, p_2)$ depends on the difference in right ascensions and the cosine of the declination of both points. For this reason, it is convenient to work with the quantity $\alpha^* = \alpha.cos(\delta)$ instead of α directly.

Catalogs usually give the right ascension α, declination δ and *visual magnitude* m (a measure of brightness) of each star. These parameters are modeled as a multivariate normal distribution. However, in several catalogs, each parameter is considered independent from each other. Therefore, for a given star we have $\overline{\alpha^*} \sim \mathcal{N}(\alpha^*, \sigma^2_{\alpha^*})$, $\overline{\delta} \sim \mathcal{N}(\delta, \sigma^2_{\delta})$, $\overline{m} \sim \mathcal{N}(m, \sigma^2_{m})$, where α^*, δ and m are the expected values of the parameters and σ_{α^*}, σ_δ and σ_m its standard errors.

Positions provided in a catalog are valid for a certain *epoch*, which is a specific moment in time. However, there exist transformations for translating positions from one epoch to other such as precession and nutation. In addition, stars have its own apparent motion across the sky denominated *proper motion*. Some catalogs also provide additional coefficients for computing the correction in proper motion. These coefficients have its own standard errors. Therefore, it is possible to compute the positions and its uncertainties of a star for a new epoch by means of the mentioned transformations and the propagation of the error [9]. This is the case of the catalog PPMX [10] where position for epoch $J2000.0$, brightness, proper motions and its uncertainties are available, among others parameters.

Naturally, older catalogs handle less information. For instance, the Cordoba Durchmusterung (CD) does not report standard errors for each star, but a mean standard error over several stars from the same region of the sky [11], e.g. for stars whose declinations are between $-22°$ and $-32°$ we have $\sigma_{\alpha^*} = 9.3$ arcsec and $\sigma_\delta = 20.5$ arcsec.

Some extra parameters $(\rho_{max}, d_{sep}, \sigma_{d_{sep}}, \sigma_{mag}, p_\emptyset, p_{sgl})$ must be determined before performing the cross-identification. Therefore, the input of our problem

[9] Details of these transformations are treated in J. Kovalevsky and P. K. Seidelmann, *Fundamentals of Astrometry*, Cambridge University Press, UK, 2004.

[10] See S. Roeser, E. Schilbach, H. Schwan, N. V. Kharchenko, A. E. Piskunov and R.-D. Scholz, *PPM-Extended (PPMX), a catalogue of positions and proper motions*, Astron. Astrophys. **488** (2008), 401–408.

[11] See pages XXIX-XXX of J. M. Thome, *Cordoba Durchmusterung (-22° to -32°)*, Resultados del Observatorio Nacional Argentino **16** (1892).

consists of catalogs A, B and these extra parameters. They will be introduced thoughout this section.

Treatment of single stars. Let $a \in A \backslash A_2$ and $b \in B$. Observe that, if a and b are far from each other, it makes little sense that both represent the same star. Usually, a criterion based on the angular distance between them can be used to keep those "close" pairs. Consider a candidate for a to every star $b \in B$ such that $\theta(a,b) < \rho_{max}$ where ρ_{max} is a given threshold. Hence, let us define

$$P_a = \{\emptyset\} \cup \{\{b\} : \theta(a,b) < \rho_{max},\ b \in B\}.$$

Note that the set $\emptyset$ is added to P_a since it could happen that a star of catalog A has no counterpart in B.

Let $a \in A \backslash A_2$ and $\{b\} \in P_a$, with its corresponding values α_a^*, δ_a, m_a, $\sigma_{\alpha_a^*}$, σ_{δ_a}, σ_{m_a} and α_b^*, δ_b, m_b, $\sigma_{\alpha_b^*}$, σ_{δ_b}, σ_{m_b} respectively. A way to measure the probability that a and b are the same star is through the distribution of the 3-dimensional random vector $(\overline{\alpha_a^*} - \overline{\alpha_b^*}, \overline{\delta_a} - \overline{\delta_b}, \overline{m_a} - \overline{m_b})$, which is known that it behaves as a multivariate normal distribution whose probability density function is

$$PDF_1(x,y,z;a,b) = pdf_{dif}(x;\alpha^*,a,b).pdf_{dif}(x;\delta,a,b).pdf_{dif}(x;m,a,b)$$

where

$$pdf_{dif}(x;\tau,a,b) = pdf(x;\tau_a - \tau_b, \sigma_{\tau_a}^2 + \sigma_{\tau_b}^2), \qquad \tau \in \{\alpha^*, \delta, m\}$$

and $pdf(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ is the well known probability density function of $\mathcal{N}(\mu,\sigma^2)$. Now, define the probability that a corresponds to some $j \in P_a$ as follows:

$$p(a \to j) = \begin{cases} p_\emptyset, & j = \emptyset, \\ (1 - p_\emptyset).\dfrac{PDF_1(0,0,0;a,b)}{\sum_{\{b'\} \in P_a} PDF_1(0,0,0;a,b')}, & j = \{b\}, \end{cases}$$

where $p_\emptyset$ is an estimate of the probability that a star from A does not have counterpart in B (usually very low).

This treatment generalizes the criterion based on the "normalized distance" [12] for assigning stars from A to B, that is to assign $a \in A$ and $b \in B$

[12] See, for instance, W. Sutherland and W. Saunders, *On the likelihood ratio for source*

in a way that

$$ND(a,b) \doteq \sqrt{\left((\alpha_a^* - \alpha_b^*)/\sigma_{\alpha^*}\right)^2 + \left((\delta_a - \delta_b)/\sigma_\delta\right)^2}$$

is minimized, where σ_{α^*} and σ_δ are the lengths of the axes of the error ellipse:

Lemma 3.3 *If $\mathscr{K} = 1$, $|B| \geq |A|$, $p_\emptyset \in \mathbb{R}_+$ is almost zero, $\rho_{max} = 180°$, $\sigma_{\alpha^*} = \sigma^2_{\alpha_a^*} + \sigma^2_{\alpha_b^*}$ and $\sigma_\delta = \sigma^2_{\delta_a} + \sigma^2_{\delta_b}$ for all $a \in A$ and $b \in B$, visual magnitudes are not considered (i.e. $m_a = m_b = 0$ and $\sigma_{m_a} = \sigma_{m_b} = 1$ for all $a \in A$ and $b \in B$) and f^* is an optimal assignment then f^* is a minimum of $ND(f) \doteq \sum_{a\in A} ND(a, f(a))$.*

Proof. Note that, for each $a \in A$, $P_a = \{\emptyset\} \cup \{\{b\} : b \in B\}$ since $\theta(a,b) < \rho_{max}$ for all $a \in A$ and $b \in B$. Let $\beta = (1 - p_\emptyset)/\sum_{b'\in B} PDF_1(0,0,0;a,b')$. Then,

$$p(a \to \{b\}) = \beta PDF_1(0,0,0;a,b) = \beta \frac{1}{\sigma_{\alpha^*}\sqrt{2\pi}} e^{-\frac{(\alpha_a^* - \alpha_b^*)^2}{2\sigma^2_{\alpha^*}}} \frac{1}{\sigma_\delta\sqrt{2\pi}} e^{-\frac{(\delta_a - \delta_b)^2}{2\sigma^2_\delta}} \frac{1}{\sqrt{2\pi}}.$$

The hypothesis asserts that $p_\emptyset$ is small enough to satisfy $p(a \to \{b\}) > p_\emptyset$ for all $b \in B$. Let f be a valid assignment. W.l.o.g., suppose that $f(a) \neq \emptyset$ for all $a \in A$. Then,

$$\begin{aligned} w(f) = \sum_{a\in A} w_{af(a)} = -\sum_{a\in A} ln(p(a \to f(a)) = \\ -\frac{\beta|A|}{\sigma_{\alpha^*}\sigma_\delta(\sqrt{2\pi})^3} - \sum_{a\in A}\left(-\frac{(\alpha_a^* - \alpha^*_{f(a)})^2}{2\sigma^2_{\alpha^*}} - \frac{(\delta_a - \delta_{f(a)})^2}{2\sigma^2_\delta}\right) = \\ -\frac{\beta|A|}{\sigma_{\alpha^*}\sigma_\delta(\sqrt{2\pi})^3} + \frac{1}{2}\sum_{a\in A}\left(\frac{(\alpha_a^* - \alpha^*_{f(a)})^2}{\sigma^2_{\alpha^*}} + \frac{(\delta_a - \delta_{f(a)})^2}{\sigma^2_\delta}\right) \end{aligned}$$

If f^* is an assignment that minimizes the function w, it also minimizes ND.□

Treatment of double stars. Let $a \in A_2$ and $\{b_1, b_2\} \in P_a$ (as in the case of single stars, P_a must be obtained with an astrometric criterion such as the one presented in [2]), with its corresponding values α_a^*, δ_a, m_a, $\sigma_{\alpha_a^*}$, σ_{δ_a}, σ_{m_a}, $\alpha^*_{b_1}$, δ_{b_1}, m_{b_1}, $\sigma_{\alpha^*_{b_1}}$, $\sigma_{\delta_{b_1}}$, $\sigma_{m_{b_1}}$, $\alpha^*_{b_2}$, δ_{b_2}, m_{b_2}, $\sigma_{\alpha^*_{b_2}}$, $\sigma_{\delta_{b_2}}$, $\sigma_{m_{b_2}}$, and such that $m_{b_1} < m_{b_2}$, i.e. b_1 is brighter than b_2. The way to compute the probability that

identification, Mon. Not. R. Astron. Soc. **259** (1992), 413–420.

a corresponds to a candidate pair $\{b_1, b_2\}$ highly depends on what is meant by "double star" in catalog A. In our approach, two features are considered: the angular separation $\theta(b_1, b_2)$ and the difference in magnitude $m_{b_1} - m_{b_2}$. Let us assume that both features are independent and normally distributed, the first one as $\mathcal{N}(d_{sep}, \sigma_{d_{sep}})$ and the second one $\mathcal{N}(0, \sigma_{mag})$, where d_{sep}, $\sigma_{d_{sep}}$ and σ_{mag} are extra parameters. Then, the probability of a pair $\{b_1, b_2\}$ is a "candidate" is given by a bidimensional random vector whose first component is the difference between $\theta(b_1, b_2)$ and d_{sep}, and the second component is the difference in magnitude $m_{b_1} - m_{b_2}$. Now, the probability that a corresponds to $\{b_1, b_2\}$ is given by the probability that a and b_1 are the same star and $\{b_1, b_2\}$ is a candidate pair. The following formula defines the probability density function of a 5-dimensional random vector that comprises all together:

$$PDF_2(x.y, z, w, t; a, b_1, b_2) = PDF_1(x, y, z; a, b_1).$$
$$pdf(w; \theta(b_1, b_2) - d_{sep}, \sigma^2_{\theta(b_1,b_2)} + \sigma^2_{d_{sep}}).pdf(t; m_{b_1} - m_{b_2}, \sigma^2_{m_{b_1}} + \sigma^2_{m_{b_2}} + \sigma^2_{mag})$$

where $\theta(b_1, b_2)$ and $\sigma_{\theta(b_1,b_2)}$ can be computed from position and standard errors of b_1 and b_2. Now, define the probability that a corresponds to some $j \in P_a$ as follows:

$$p(a \to j) = \begin{cases} p_\emptyset, & j = \emptyset, \\ (1 - p_\emptyset).p_{sgl}.\dfrac{PDF_1(0, 0, 0; a, b)}{\sum_{\{b'\} \in P_a} PDF_1(0, 0, 0; a, b')}, & j = \{b\}, \\ (1 - p_\emptyset).(1 - p_{sgl}).\dfrac{PDF_2(0, 0, 0, 0, 0; a, b_1, b_2)}{\sum_{\{b'_1, b'_2\} \in P_a} PDF_2(0.0, 0, 0, 0; a, b'_1, b'_2)}, & j = \{b_1, b_2\}, \end{cases}$$

where p_{sgl} is an estimate of the probability that a double star from A may be assigned to some single star in B.

Preprocessing catalogs. The resolution given in Section 2 consists of the cross-identification performed between two known stellar catalogs. The former one is a part of CD (catalog I/114 of VizieR astronomical database) consisting of 52692 stars whose declinations are between $-22°$ and $-25°$ for epoch $B1875.0$. The reason for taking these subset of stars is that the information about double stars, i.e. the set A_2, is only available in printed form and must be entered by hand. In our case, 571 stars were transcribed, corresponding to the first 177 pages of the printed catalog.

The other catalog is a part of PPMX (catalog I/312 of VizieR) with 130664 stars which cover the sky region of the former one.

The preprocessing of both catalogs is essentially the same as in [2]. Some

stars from CD have been deliberately removed due to the following causes: 1) variable star; 2) cumulus; 3) a star appearing in PPM (catalogs I/193, I/206 and I/208 of VizieR) and whose position in PPM differs from CD in more than 2 arcmin for epoch $B1875.0$ or whose magnitude differs from CD in more than 1.5. Some other entries in catalog CD has been altered because of typo errors [2]. After this process, there are 52313 stars left (where 568 are doubles).

Data from PPMX catalog have been preprocessed as follows. Visual magnitudes have been converted to the magnitude scale used by CD: $m_{CD} = -0.01335368m^2 + 1.076636m + 0.2249828$ where m is the Johnson V magnitude reported in PPMX and m_{CD} is the target magnitude. These coefficients have been obtained through a quadratic fit explained in [2]. Positions have been translated to the epoch of CD. In addition, the column of visual magnitude (specifically, Johnson V) for several entries of PPMX is empty so it has been filled with magnitudes from catalog APASS-DR9 (catalog II/336 of VizieR). After this process, stars with magnitude greater than 13.5 have been discarded, leaving 83397 stars.

A preliminary cross-identification between CD and PPMX has been performed via the X-Match Service in order to generate the sets of candidate stars P_a faster. The parameters and standard errors have been set as follows:

- $\rho_{max} = 2$ arcmin (the maximum allowed by X-Match)
- $d_{sep} = 34.9$ arcsec [2]
- $\sigma_{d_{sep}} = 13.65$ arcsec [2]
- $\sigma_{mag} = 0.915$ [2]
- $p_\emptyset = 10^{-10}$
- $p_{sgl} = 10^{-4}$
- $\sigma^2_{\alpha^*_a} + \sigma^2_{\alpha^*_b} = (10.15 \text{ arcsec})^2$, $\sigma^2_{\delta_a} + \sigma^2_{\delta_b} = (22.74 \text{ arcsec})^2$, $\sigma^2_{m_a} + \sigma^2_{m_b} = 0.2759^2$ for all $a \in A$ and $b \in B$ [2]
- $\sigma^2_{\theta(b_1,b_2)} = 0$, $\sigma^2_{m_{b_1}} + \sigma^2_{m_{b_2}} = 0$ for all $b_1, b_2 \in B$

The dataset [3] contains the new CD catalog with the cross-identification (`new_cd.txt`), its format (`new_format.txt`), the source code as well as other auxiliary files. In Figure 6 a picture of the whole process is displayed.

Acknowledgements. I would like to thank María Julia Severín and the people mentioned in the Acknowledgements of [2] who help me to enter the set of double stars, among other data.

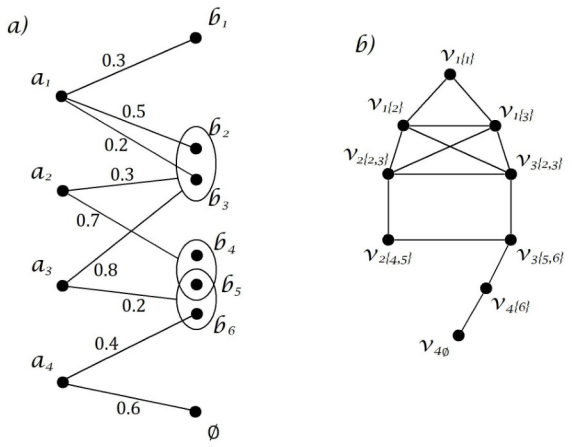

Fig. 3. Example of 2-Matching Problem: a) instance, b) G

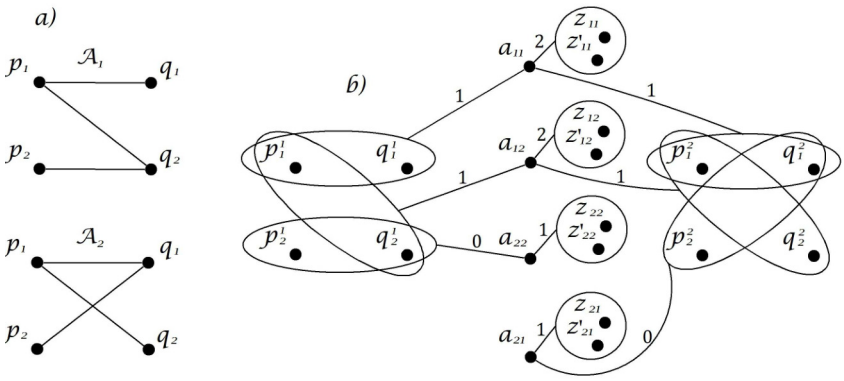

Fig. 4. Example of reduction: a) DM, b) 2-MDPAW

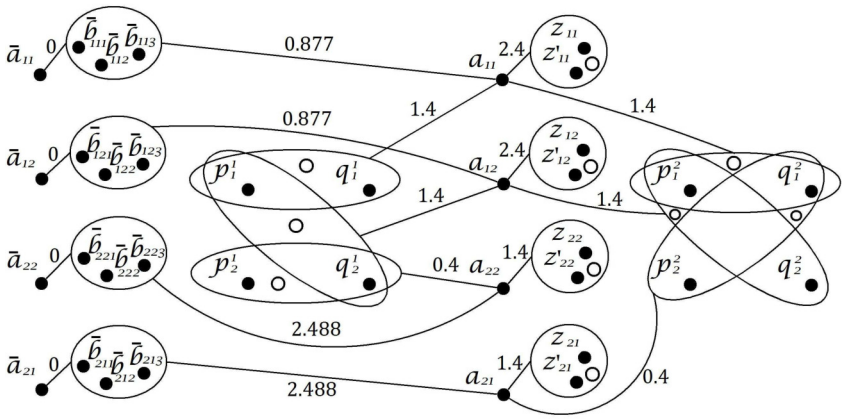

Fig. 5. Example of reduction to 3-MDP

Number of stars of catalog A (n_A)	52313
Number of stars of catalog B (n_B)	83397
Double stars present in catalog A ($\|A_2\|$)	568
Largest cardinal of P_a	34
Number of components of G:	
• Unique star	39383
• Only single stars	5628
• Multiple stars	111
Largest cardinal of A' found in components:	
• Only single stars	34
• Multiple stars	7
Statistics of thc solution:	
• Unassigned stars	245
• Single stars assigned	51502
• Double stars assigned	483
• Double stars assigned to a single one in B	83

Table 1
Highlights about the optimization

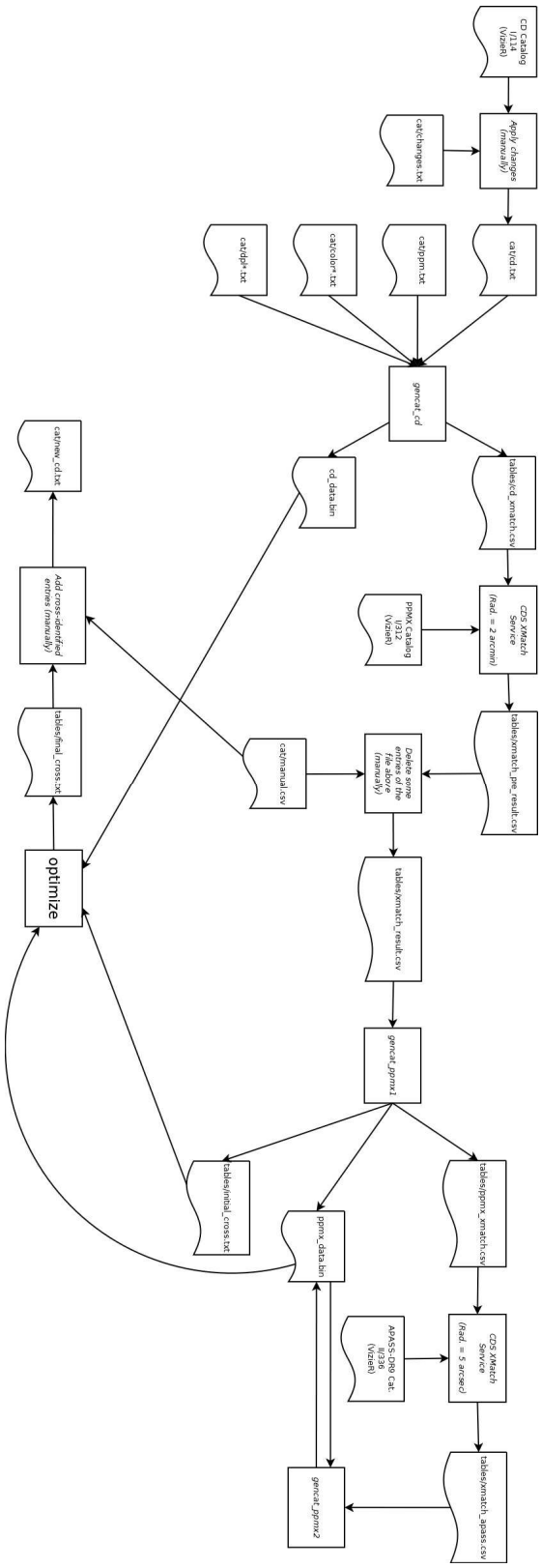

Fig. 6. Diagram