Research Article

Sergei Sapozhnikov* and Dana Kovaleva

Application of clustering algorithm to wide stellar pairs for unsupervised search of parts of disrupting clusters

<https://doi.org/10.1515/astro-2021-0025>

Received Oct 31, 2021; accepted Dec 10, 2021

Abstract: We introduce the application of the clustering algorithm to the preliminary compiled list of probable wide pairs of co-moving stars. The main aim of such development is a possibility of unsupervised blind search of coeval loose stellar structures over significant regions in space. Using Gaia EDR3 data, we investigated the application of the method to nearby region hosting recently discovered loose structures – tidal tails of Coma Ber star cluster and a nearby stellar group named Group X. We compare the results of straightforward clustering of stellar data with results of using our method with varying parameters. We then compare the results of our method to the recent results of the two groups of authors who independently discovered the discussed structures. We find parameters allowing the method to find the full scope of distributed stellar groups without preliminary knowledge of their characteristics. It decreases the risk of false positive clustering and improves the ability to discover loose stellar groups, in comparison with the application of clustering algorithm to the individual stars. Further we obtain a refined dataset of probable members of both stellar groups and independently obtain their ages (700 ± 70 Myr and 450 ± 100 Myr) and space velocities $((U,V,W) = (8.63 \pm 0.13, 6.63 \pm 0.20, 6.65 \pm 0.95)$ km/s for Coma Ber star cluster, and $(U,V,W) = (7.70 \pm 0.12, 3.27 \pm 0.45, 5.69 \pm 0.80)$ km/s for Group X). Our results are in very good agreement with those of previous investigators.

Keywords: open clusters, tidal tails, stellar streams, Gaia

1 Introduction

Star clusters provide important clues for fundamental problems of structure, history, and evolution of the Galaxy. Astrometric space mission Gaia (Brown *et al.* 2016) provides high accuracy data on positions, proper motions, and parallaxes of stars with previously unattainable precision and scale (> 1.5 billion sources) (Brown *et al.* 2021). This, in particular, enabled new achievements in the discovery of star clusters and obtaining their membership (Cantat-Gaudin *et al.* 2018; Castro-Ginard *et al.* 2018, 2020) as well as the discovery of numerous loose stellar groups originated from different stages of disruption of star clusters (Jerabkova *et al.* 2021a; Meingast *et al.* 2021; Röser and Schilbach 2020; Ratzenböck *et al.* 2020; Tang *et al.* 2019). While stars in such structures might be unbound at present time, they do

share a common genesis. This makes them important for an understanding of Galactic structure, dynamical evolution, as well as the formation of a population of the Galaxy stellar field.

For the census of star clusters, unsupervised clustering algorithms proved to be quite successful, though for different goals they may be different (see, e.g., discussion in Hunt and Reffert (2021)). Loose stellar structures related to existing clusters (tidal tails, stellar coronae, filaments) (Jerabkova *et al.* 2021a; Meingast *et al.* 2021; Tang *et al.* 2019; Zhong *et al.* 2019); moving groups and stellar streams (Röser and Schilbach 2020; Ratzenböck *et al.* 2020) are more tricky for discovery with clustering algorithms. They may barely stand out over the background, have a complicated geometric form, or have a significant spatial extent which results in variations of proper motions.

Authors explore various approaches to locate such structures. One approach often seen is to pick an already known stellar cluster or group, and search in phase space in Gaia data around it for extended structures kinematically related to it, e.g. Jerabkova *et al.* (2021b); Röser and Schilbach (2019); Zhong *et al.* (2019); Pang *et al.* (2020). These methods are usually too slow for large-scale surveys.

Corresponding Author: Sergei Sapozhnikov: Institute of Astronomy, Russian Academy of Sciences, Russian Federation; Email: thestriks@gmail.com

Dana Kovaleva: Institute of Astronomy, Russian Academy of Sciences, Russian Federation

Another approach involves the application of clustering methods like DBSCAN/HDBSCAN to a stellar population in general (Hunt and Reffert 2021). Search for clusters in phase space (coordinates + velocities) allows finding new structures unrelated to known ones since this method does not rely on finding the starting point first. The chance of discovery of an extended, loose group of stars depends on its characteristics and on the parameters of used methods. All-sky surveys for ultra-wide binary stars in Gaia data (El-Badry *et al.* 2019; Jiménez-Esteban *et al.* 2019; Sapozhnikov *et al.* 2020) may also be considered in this spirit.

We combine the approaches: we use our algorithm designed for search of ultrawide binaries described in Sapozhnikov *et al.* (2020) to look for comoving stellar pairs and use the DBSCAN (Ester *et al.* 1996) clustering algorithm to this preliminary list of stellar pairs (instead of using it on star catalogue directly). The selection of DBSCAN is based on the work by Hunt and Reffert (2021) highlighting DBSCAN and HDBSCAN (Campello *et al.* 2013) as the best algorithms for clustering such data. We choose DBSCAN over HDBSCAN due to the possibility to explicitly set the distance between points in phase space as a clustering parameter (so-called ε -parameter in DBSCAN). A preliminary compilation of the list of comoving pairs would allow us to constrain the parameters of relative movement of stars before applying the clustering algorithms. Limits on parameters are selected so that random pair picked from an extended structure has a higher chance to pass them (compared to a random pair from stellar background). This increases the contrast with which pairs constituting extended stellar structures appear in the phase space, which in turn increases the algorithm sensitivity. This advantage over direct clustering of stars is shown in more detail later in 3.2. The main goal is the development of an unsupervised algorithm of the search for extended stellar structures which is sufficiently fast to be applied blindly and systematically over the celestial sphere, and sufficiently sensitive to find such structures.

We present the description of the method and its testing with model data and over the region of the sky including extended stellar groups previously identified by other authors (Fürnkranz *et al.* 2019; Tang *et al.* 2019). Section 2 describes the data we investigate. Section 3.1 describes the principles behind creating the preliminary list of pairs. Section 3.2 describes a model distribution of stars and the application of the clustering algorithm to mock data. In section 4 we apply the method to real data in the region hosting Coma Ber star cluster, and in section 5 the results and plans are discussed.

2 Data

To test how well our method discovers loose stellar structures, and how altering the clustering algorithm parameters and pair parameter restrictions affect this process, we apply the method to the region of the sky where the Coma Ber star cluster (Melotte 111) is located.

The reason to select this region is the recent discovery of elongated tidal structures around the Coma Ber cluster, as well as non-fiducial loose stellar group in the vicinity with Gaia data. These discoveries were made almost simultaneously and independently by the two groups (Fürnkranz *et al.* 2019; Tang *et al.* 2019) using different methods applied to Gaia DR2 data, and obtaining similar, but slightly different in respect to membership, results.

Fürnkranz *et al.* (2019) discovered overdensities in velocity space with wavelet decomposition method, apply filters, and then use DBSCAN to the preliminary selection of 11294 sources to locate members of extended groups. They refer 214 stars to possible members of the Coma Ber cluster, and 177 stars — to its neighboring group. Tang *et al.* (2019) primarily search around Coma Ber for tidal tails, applying cuts in proper motion to decrease the number of sources to be considered to 5494, and then use a method based on the self-organizing map, StarGo (Yuan *et al.* 2018) to identify relevant structures. This work marks 197 stars as possible members of the Coma Ber cluster, and 218 — as members of Group X. The datasets selected in these two papers overlap but do not coincide: there are 174 common stars in selections of sources referred to Coma Ber cluster, and 156 common stars referred to Group X. We continue the discussion in paragraph 4.

The newly discovered group, named by Tang *et al.* (2019) “Group X”, differs from Coma Ber cluster by age. Fürnkranz *et al.* (2019) provide approximate ages equal to 400 Myr and 700 Myr, respectively. Tang *et al.* (2019) estimate age of Coma Ber cluster between 700-800 Myr, and ≈ 400 Myr for Group X. Along with small relative errors in parallaxes, due to closeness of the two groups to the Sun (85 and ~ 100 pc, respectively), this helps us use colour-absolute magnitude diagrams to check possible contamination of selected members with field stars.

The discussed region and two hosted groups of stars provide an interesting challenge due to variety of different structures: relatively dense core of Coma Ber, its loose tidal tails, and elongated irregular structure of Group X.

We select all Gaia EDR3 sources satisfying conditions: (circle of 40 – degree radius around $l = 110$, $b = 80$ in Galactic coordinates; $8 < \varpi < 17$ mas). The resulting list includes 56188 stars. We will refer to it as Sample I.

3 Method

3.1 Catalogue of wide pairs of stars

Assembling a preliminary pairs catalogue involves picking several parameters of stellar pairs and putting constraints on them to differ from random contamination. Search for pairs then is done by filtering lists of all possible pairs in the catalogue to fit the constraints. These constraints are based on ranges of these parameters expected for pairs of stars comprising extended stellar structures. For Gaia EDR3, the obvious selection of parameters we may lay on is α , δ , ϖ , μ_α^* , μ_δ where α , δ are equatorial coordinates of stars, ϖ is parallax, and μ_α^* , μ_δ are components of proper motion (hereafter $\mu_\alpha^* = \mu_\alpha \cdot \cos \delta$). Radial velocities are not considered in the process of selection, since they are present only for a small fraction of all Gaia EDR3 sources and provide a potential tool for independent control of the results.

For the selection of constraints on pairs properties, we guide ourselves with theoretical and model results from Kamdar *et al.* (2019). Selected constraints are what we roughly expect from a stellar pair of neighbours in the same extended stellar structure. We investigate how varying the maximum distances and velocity difference between the components of pair affect the resulting list of pairs, remaining within limits where the majority of pairs are genuine and the probability of random contamination is low. This covers tentative separations between the two stars $\rho_0 \leq 12$ pc and differences in velocities $\Delta V \leq 2$ km/s. We apply somewhat loosened criteria to avoid missing candidates for pairs due to observational errors. The further procedures are sufficient to filter out random contamination (see paragraph 3.2). The main purpose of the creation of a catalogue of pairs is to increase the contrast of comoving objects against the stellar background. Pairs within comoving objects become relatively more prevalent thanks to taking stars within a group into account multiple times due to it forming a number of pairs. Simultaneously, stars without any comoving companions in their vicinity are excluded, thus decreasing background contamination, as would be shown below. These structures with higher contrast make for more easily discernible targets for the DBSCAN algorithm.

For every two stars, we compile two criteria for spatial separation, due to uncertainty of position on the celestial sphere being negligible compared to distance uncertainties, and one for the difference of relative motion of stars.

Among two criteria determining whether stars are being relatively spatially close to one another, the first one

limits projected separation between stars to ρ_0 parsec (we consider results with a selection of ρ_0 , see Table 1). This limit also allows us to significantly optimise the pairs catalogue assembly process by subdividing search region into smaller regions on the celestial sphere, sizes of which are determined by maximum possible coordinate separation of a pair and then searching for pairs just within these sub-regions and in neighbouring ones. This allows to decrease computation complexity from $C * N^2$ to $C * N^2 / M$, where N is the number of stars and M is the number of regions ($M = 1537$ in case of $\rho = 1$ pc), which meant a significant gain in computation time. This optimisation is reviewed in more detail in Sapozhnikov *et al.* (2020).

The projected separation in pc between two stars is calculated as

$$\rho_{12} = \frac{2}{\varpi_p} \sqrt{\sin^2 \left[\frac{\delta_2 - \delta_1}{2} \right] + \cos \delta_1 \cos \delta_2 \sin^2 \left[\frac{\alpha_2 - \alpha_1}{2} \right]}, \quad (1)$$

where $\varpi_p = \frac{\varpi_1 + \varpi_2}{2}$, mean parallax of two stars in question in arcsec.

Error in ρ_{12} linearly depends on relative error in parallax.

Relative errors in parallax are significantly larger than relative errors of position in the tangential plane. Inferring distances to stars from parallaxes is not a straightforward task (Bailer-Jones 2015), and proper Bayesian inference of distance between every two stars would have been unnecessary time-consuming for the intermediate task of filtration of possible pairs. “Naive” estimate of the distance to every star as an inverse of parallax leads to a very uncertain estimate of the line-of-sight distance between two stars, taking into account that relative errors of parallax increase for distant stars, and symmetric errors of parallax values lead to asymmetric errors in distance determination. This motivates us to compile a criterion for the line-of-sight distance between stars based on parallaxes and their errors ($\varpi_1 \pm \epsilon_{\varpi_1}$, $\varpi_2 \pm \epsilon_{\varpi_2}$) instead of introducing a constraint on apparent measured distances difference.

We decide to impose limits on parameter further referred as “parallax consistency”:

$$\varpi_{cons} = 3(\epsilon_{\varpi_1} + \epsilon_{\varpi_2}) - |(\varpi_1 - \varpi_2)|, \quad (2)$$

where ϖ_1 , ϖ_2 are parallaxes of components, and ϵ_{ϖ_1} , ϵ_{ϖ_2} are their errors. Negative values of ϖ_{cons} indicate that parallaxes are too different for stars to be close even if we allow overlap within 3 standard deviations for both components. Highly positive values indicate that errors are too large to tell with certainty that stars are close together. Values of ϖ_{cons} that are positive, but close to zero indicate either that stars are close together with relative certainty, or that the

Figure 1. Scatterplots illustrating the parallax consistency $w_{cons} = 3(\epsilon_{w_1} + \epsilon_{w_2}) - |w_1 - w_2|$ selection process on a small sample of 65000 pairs formed from real stars random sample with 8pc maximum separation and no limits on parallax consistency. On both scatterplots region of parallax consistency from -0.1 to 1 is shown by vertical dashed lines.

On the left: relation between parallax consistency and modelled probability for pair components to be within 5 pc of one another along line of sight (in percent). Scale for probability is logarithmic. On the right: Relation between consistency of parallaxes within errors and relative motion of pair components

sum of errors happened to be close to parallax difference (even if it is large).

To better understand which values of w_{cons} suit our needs, we investigate a model simulating probability density distribution of the radial distance between stars by modelling probability distribution function of distance to Sun for both stars, based on their given parallaxes and their errors, and then integrating over the limit $(-5 \text{ pc}, +5 \text{ pc})$ for each pair to determine the probability of them being within 5 pc of one another along line-of-sight (P_5). We see that these two metrics do somewhat correlate (see Figure 1): most of the pairs with negative parallax consistencies have $P_5 < 0.2\%$, and high P_5 is mostly associated with positive but not very high w_{cons} . We decided to not impose limits on P_5 directly, but we do use a limit on parallax consistency. The lower limit is selected to be -0.1 mas (we exclude the pairs where parallax difference is larger than 3 times errors, plus 0.1 mas margin). Deciding on the upper limit is not so obvious, we decide by checking the dependency of parallax consistency vs. relative motion of components in pair, see Figure 1.

We define “relative motion” for a pair of stars as

$$V_{\Delta}[\text{km/s}] = \frac{\kappa}{10^3 w_p} \sqrt{(\Delta_{\mu RA})^2 + (\Delta_{\mu Dec})^2} \quad (3)$$

Here, $\Delta_{\mu RA}$ and $\Delta_{\mu Dec}$ are proper motion differences between two stars in mas/yr , and w_p is the mean parallax between two stars in arcsec , and $\kappa = 4.74$ is the transformation factor from 1 mas/yr at 1 kpc to 1 km/s .

We note that the majority of pairs with small relative motion concentrate within 1 mas of parallax consistency, so we select the upper limit of parallax consistency to be 1 mas .

We create a pairs catalogue with following restrictions on parameters:

- $V_{\Delta} < 2 \text{ km/s}$,
- $\rho_{12} < 8 \text{ pc}$,
- $-0.1 \text{ mas} < w_{cons} < 1 \text{ mas}$.

We will investigate other options on V_{Δ} and ρ_{12} later, but first, we will apply the clustering algorithm to this particular set of pairs, and compare the results with the mock catalogue we will describe in the next section.

3.2 Stellar background model and parameters of clustering algorithm

Construction of catalogue of possible wide pairs allows defining overdensities in the space [distance between two stars, their relative velocity]. After that, we may apply a

clustering algorithm to find whether these overdensities do form any structures.

There are two traits we want our method to have. The first is successfully discovering structures that do exist (avoiding false negatives). The second is how well it avoids mistaking random contamination for a real object (avoiding false positives).

The stellar background model aim is to produce a mock catalogue of single stars that would be known to not contain any clusters, streams, or other stellar structures. We apply the process of construction of catalogue of preliminary pairs to this catalogue and perform clustering on them, to check whether our algorithm avoids false positive results. Since whether false positives will appear will depend on typical errors in observational parameters (primarily, parallaxes and proper motions), limitations of the method should be checked depending on the range of distances where it is going to be applied.

Initially, the principles of creation of mock catalogue were described in Sapozhnikov and Kovaleva (2020). Here we adopt them for the present task. The following model is used: stars are first generated uniformly within the same region of the sky as the real sample, but in a broader range of distances: 35 – 260 pc instead of 70 – 130 pc. To simulate the effect of parallax errors on measuring actual parallaxes, each star is assigned a “real” parallax in mas ($1000./\text{distance}$) and a parallax measurement error. The distribution of parallax errors was made to mimic that of the real sample. “Measured” star parallaxes are obtained by modifying their “real” parallaxes by a random number drawn from the normal distribution with the standard deviation equal to parallax error. After that the dataset is cropped to 70 – 130 pc of “measured” distance. Stars get random galactocentric velocity components (U , V , W) with dispersions of 35, 30, and 20 km/s respectively to simulate the actual movement of stars in solar neighbourhood (Schönrich *et al.* 2010). The density of stars in the model is roughly the same as the density of stars for real EDR3 data ($\sim 70\,000$ stars in volume in question). We apply the algorithm to find stellar pairs to both real data and the model with the same parameters and constraints we described in the previous section. This results in 17016 pairs in real data 4705 and pairs in model data. Comparison of density distribution on the sky for real and model pairs can be seen in Figure 2. It can be seen how pairs from real data reveal very distinct features that are not present in mostly uniformly distributed model pairs.

We apply a DBSCAN clustering algorithm to both real and model data. Clustering is done in 4-parameter space (two for positions on the sky and two for proper motions). The radial distance is not considered due to parallax errors stretching any stellar structure in line-of-sight direction too

Figure 2. Scatterplots showing the sky positions of pairs from real (left) and background model (right) data. More densely populated areas are closer to purple color. Generated pairs are notably more scarce

much, restriction on w_{cons} in the pair composition phase is considered to be sufficient to take radial distances into account. DBSCAN has two available parameters for clustering. First is the so-called “ ϵ ” value: DBSCAN defines that two samples in the dataset are in the “neighbourhood” of each other if the distance in 4-parameter space between them is less than ϵ . Cluster is formed when some minimum amount of samples fall within one neighbourhood: this

minimal amount is the second DBSCAN parameter called “min_samples”. For clustering of pairs, choosing the right pair of parameters is not obvious, so we investigate by applying the clustering algorithm with varying values of (ϵ , min_samples) to four samples: pairs created from real stars, pairs created from the model, real stars, and model stars. We observe the differing number of clusters labelled in data

by the DBSCAN algorithm depending on parameters, this relation can be seen in Figure 3.

We expect that for the clustering process to not produce false positives, we should pick the parameters that do not lead to the “discovery” of clusters in mock data that is intentionally devoid of any stellar structures. It can be seen in Figure 3 how parameters area where real data do produce

Figure 3. Dependency of amount of clusters labelled by DBSCAN on the clustering parameters for real (left) and model (right) data, for clustering of pairs (top) and stars directly (bottom). Amount of clusters labelled is shown with color gradient, green-to-white isolines designate areas with >1 , >2 , >4 and >10 clusters. Parameter value areas which produce clustering in model data are treated as the ones prone to creating false positives. It is therefore concluded that parameters of clustering should fit below area of false-positives, but in the region where real data does produce clusters. It can be seen that this suitable region is larger for the pairs data (compared to direct clustering of stars). The real pairs clustering parameters elected for further use are marked by black cross on the corresponding (top-left) image

Figure 4. Scatterplots showing the sky positions in (l,b) galactic coordinates (left) and proper motions (right) of stars from clustered real data, coloured in according to which cluster they were allocated to. Grey smaller dots correspond to the stars which were present at least in one pair, but were not attributed to clusters by DBSCAN algorithm

clusters, but mock data do not is larger for the pairs (compared to direct clustering of stars). These figures show how a range of parameters can be a valid choice. To make the choice, we guide ourselves with the following judgment: we expect the average projected distance between pairs in an extended structure to be less than the distance be-

tween stars in it (since one star can be present in more than one pair). We decide that a reasonable maximum of spatial separation between pairs would be around 60% of maximum spatial separation between stars, which is ~ 5 pc. Spatial separation ~ 5 pc for the average distance to stars in our sample would mean an $\varepsilon \sim 6$. As for the minimal samples parameter, the relation between a number of stars and a number of pairs within a structure depends on how stretched out the structure is. If we define M as the number of stars and N as the number of pairs, the structure stretched into a thin line has $N = (M-1)$ pairs. Structure dense enough so that every star in it forms a pair with every other form $M = \frac{N!}{2(N-2)!}$ pairs. We decide to go for $\text{min_samples} = 20$, which corresponds to 7-21 stars depending on the stretchiness of the structure. In conclusion, $\varepsilon = 6$ and $\text{min_samples} = 20$ adopted for further use. These parameters are within the “no false positives” region. The result of clustering with these parameters can be seen in Figure 4: here we allocate stars to clusters based on the clustering of pairs these stars are a part of. Two clusters corresponding to Coma Ber and Group X can be seen, along with several smaller clusters. We will further investigate how this situation varies with variation of the parameters of catalogue of pairs.

4 Results

We explore sample I with different approaches: straightforward application of clustering algorithm DBSCAN to the data on stars, and application of the same algorithm with the same parameters (see Section 3.2) to the set of catalogues of stellar pairs generated with different pairing parameters. We will refer to these catalogues as “Sample (i, j)” where i, j are pairing parameters: i is the limiting projected separation, in pc (in this work the following values are explored: $i = 3, 5, 8, 12$), and j is maximum allowed projected velocity difference, in km/s ($j = 1, 2$).

Table 1 summarizes information about the samples and the results of the application of a clustering algorithm to them. For each sample, we provide the number of tentative pairs in the catalogue of pairs and the number of unique stars which are components of at least one pair. We do not mark cases when all or almost all the background stars are marked as a cluster because they are obvious to be filtered out.

The clustered groups are referred either to Coma Ber or to Group X if they obviously are associated with the regions in space and in velocities covered by the respective stellar group in Fürnkranz *et al.* (2019) and/or Tang *et al.* (2019). The total number of stars related in these two publications

Table 1. Results of application of clustering algorithm to different samples

Star group /	Sample	l	(3,2)	(3,1)	(5,2)	(5,1)	(8,2)	(8,1)	(12,2)	(12,1)
	N pairs in sample	–	6112	3899	9879	6004	17016	9602	28124	14253
	N stars in sample	56188	8437	5477	10670	6239	14945	7695	21115	10119
Coma Ber	N clusters	1	1	1	2	5	2	1	1	2
	N stars	171	170	154	215	197	298	229	385	285
Group X	N clusters	1	1	2	1	1	1	1	1	2
	N stars	243	220	163	230	269	312	230	375	304
Extra	N clusters	0	0	0	0	0	1	0	4	0
	N stars	0	0	0	0	0	12	0	67	0

Figure 5. Comparison of number and distribution of sources considered as probable members of Coma Ber cluster (magenta) and of Group X (blue) obtained by different methods. The coordinate frame is Galactic (l,b). Left upper plane – samples from Fürnkranz *et al.* (2019) (crosses) vs Tang *et al.* (2019) (empty circles). The upper right plane – sample (12,1) (filled squares) vs sample l (clustering of single stars, small dots of lighter colours). The middle plane – sample (12,1) (filled squares and empty squares for small clusters, see discussion in the text) vs sample from Tang *et al.* (2019) (empty circles). Lower plane – sample (12,2) (filled dots) vs sample from Tang *et al.* (2019) (empty circles). Black, grey, yellow, green dots – clustered stars not identified with known structures with certainty

Figure 6. Stars referred to Coma Ber cluster (magenta dots) and to Group X (blue dots) along with fitted isochrones. Empty dots are for the stars filtered out because of quality of astrometric solution (not used for isochrone fitting), filled dots for astrometrically pure sources

to the Coma Ber cluster or to Group X should not be directly compared with the numbers in Table 1. Primarily, the authors of both cited papers use data of Gaia DR2, while we use Gaia EDR3. Further, they used different schemes of data filtering (we later also apply one). We think it advisable, however, to compare regions in space of parameters.

One may see from Table 1 that larger allowances for projected distance and projected velocity difference in pairing in general (though not always) lead to a larger number of stars classified as possible members of groups. The remarkable exception may be seen at samples (3,1) and (3,2) where the stricter limit on velocity difference leads to clustering of Group X into two groups. This is physically reasonable, see Figure 5: as Fürnkranz *et al.* (2019) note, Group X shows the inhomogeneous distribution of sources divided into two elongated subgroups in density distribution; Tang *et al.* (2019) state this stellar group is disrupting. This allows each subgroup to be clustered separately and effectively provides larger dispersion for velocity difference for Group X as a total. In general, a smaller limit for the difference of velocities results in similar or better results for every limiting distance between components of pair within considered limits. For better coverage of the extended loose structure, not surprisingly, larger i are more favourable, while lower values of j are advisable to avoid false positives. Processing the sample (12,1) we find two clusters that may be associated with Coma Ber, and two clusters for Group X. In both cases one of these clusters is large (271 sources for

Figure 7. Residual tangential velocities of refined samples of stars referred to Coma Ber cluster (magenta dots) and to Group X (blue dots) in respect to the mean space motion of Coma Ber (upper panel) and Group X (lower panel). The length of an arrow in degrees (in scale of the figure) corresponds to the residual velocity in km/s. Coordinate frame is Galactic (l,b)

Coma Ber and 299 for Group X) and another one is small (18 and 14 stars, respectively). The small clusters are located very close to the region occupied by a larger one and partly covered by probable members, according to other authors (Fürnkranz *et al.* 2019; Tang *et al.* 2019). We consider them to be probable parts of these two groups. At the panel of Figure 5 members of these small groups are marked by empty squares.

Samples (8,2) and especially (12,2) produce extra clusters in the regions not covered by other methods or samples (8,1) and (12,1). These small clusters may be a result of taking random contamination for a cluster. However, the application of clustering algorithm toward stellar background model does not result in false positives even at large (i,j) of tested ones. Extra clusters may be composed of smaller remnants of stellar groups, probably of the Coma Ber clus-

Figure 8. Position of stars – probable members of Coma Ber and Group X at the celestial sphere and in space. Colour marks EDR3 parallaxes of stars, darker colours represent closer stars. Coordinate frame is Galactic (l,b)

ter. Neither their colour–magnitude diagrams nor radial velocities contradict this hypothesis. Spatially these small groups, except one, are very close to Coma Ber tidal structures. This, however, requires additional investigation to prove. At the lower panel of Figure 5 the extra clusters are marked by grey, yellow, green and black (the group marked in black is about 50 pc from the centre of the Coma Ber cluster).

Our samples confirm different ages of Coma Ber and Group X. We refine the dataset obtained with sample (12,1) filtering out stars having Gaia EDR3 parameters $ruwe > 1.4$ and $parallax_over_error < 10$. This keeps 244 and 246 stars for Coma Ber and X groups, respectively. We suggest interstellar extinction to be negligible, and use isochrones by Bressan *et al.* (2012) for Gaia EDR3 magnitudes from <http://stev.oapd.inaf.it/cmd> to estimate their ages with isochrone fitting procedure laying mainly on the upper main sequence and turn-off stars, under the suggestion of a solar chemical composition (Fürnkranz *et al.* (2019) indicate that there may be a difference in metallicities of the two groups of ≈ 0.1 dex). The resulting ages are 700 ± 70 Myr and 450 ± 100 Myr for the Coma Ber cluster and Group X, respectively. Uncertainties are based on isochrone fitting applied to Monte-Carlo random sampling of colour-magnitude diagrams taking into account both random and expected systematic errors. Obtained ages perfectly agree with estimates by (Fürnkranz *et al.* 2019; Tang *et al.* 2019). Figure 6 represents stars referred to the Coma Ber cluster and to Group X along with fitted isochrones. Further, we calculate for every star of the sample value of photometric probability of membership P_{phot} (see Kharchenko *et al.* (2012) for details), in respect of age $\lg t[\text{yr}] = 8.85$ for Coma Ber cluster and $\lg t[\text{yr}] = 8.65$ for X Group, and retain in

consideration only stars having $P_{phot} > 0$. Among others, this removes from consideration falling under isochrones at Figure 6. This results in 206 and 208 stars left for consideration among the Coma Ber cluster and Group X, respectively.

Among the stars of astrometrically and photometrically refined dataset, there are 60 sources with V_r measured by Gaia DR2 belonging to Coma Ber group, and 42 sources for X Group. Based on these radial velocities, we obtain space velocities for the two investigated groups in respect to the local standard of rest (Schönrich *et al.* 2010): $(U,V,W) = (8.63 \pm 0.13, 6.63 \pm 0.20, 6.65 \pm 0.95)$ km/s for Coma Ber star cluster, and $(U,V,W) = (7.70 \pm 0.12, 3.27 \pm 0.45, 5.69 \pm 0.80)$ km/s for Group X. These values agree very well with those provided by Fürnkranz *et al.* (2019); Tang *et al.* (2019). We estimate radial velocities of these stellar groups as -0.2 ± 1.0 km/s and -6.3 ± 1.0 km/s.

Based on these values, we calculate residual tangential velocities $V_t^* - V_t^0 = 4.74 \cdot [\mu^* - \mu^0(l^*, b^*, U^0, V^0, W^0)]/\varpi^*$ for the members of Coma Ber cluster and of Group X with respect to the mean movement of each of these two stellar groups. We use the modification of convergent point method as described by Röser *et al.* (2011) to obtain for each source the difference between measured proper motion and the one it would have if sharing space motion of a certain star group. Figure 7 represents the residuals with respect to the mean movement of Coma Ber (upper panel) and with the one of Group X (lower panel). One may see the movement of the tidal structures of Coma Ber heading away from the centre of a cluster (upper panel). At the lower panel, representing view in the reference frame of Group X, it is interesting to note that in spite of the high degree of dissolution, the mean dispersion of velocities is relatively small (≈ 1.2 km/s).

Figure 8 represents the position of the two stellar groups at the celestial sphere, with an indication of distance (in terms of parallax) for every star. One may see how more remote Group X is orientated in the direction toward Coma Ber, moving closer to it.

5 Summary and conclusions

We introduce the method of application of clustering algorithm toward preliminarily selected pairs of stars with constrained differences in tangential velocities and distance between the components. The main aim of such development is a possibility of unsupervised blind search of coeval loose stellar structures over significant regions in space. Using Gaia EDR3 data, we investigated the application of the method toward nearby region hosting recently discov-

ered loose structures – tidal tails of Coma Ber star cluster and the nearby stellar group named Group X. We compare the results of the application of the method with varying parameters of the composition of stellar pairs and compare them with straightforward clustering of stellar data. We also compare our result with recent results of the two groups of authors who independently discovered the discussed structures.

The results proved that the method is able to find distributed stellar groups without preliminary knowledge of their characteristics. Experiments over the set of values of parameters of preliminary stellar pairs indicated that for such searches it is favourable to choose strict constraints on the difference of velocities (in our case, up to 1 km/s for projected velocity difference). Simultaneously, one may significantly extend the allowed distance between stars in a pair (in our experiment, the projected distance between components up to 12 pc at certain limitations over the relation between parallaxes and their errors).

We confirm that the straightforward application of the clustering algorithm toward stars of the same region has a much narrower region of allowed parameters of clustering, due to the risk of false positive results. As a result, while it discovers dense structures, the ability of the clustering algorithm to find loose stellar groups is limited.

This same clustering method but applied to pre-constructed catalogues of stellar pairs leads to re-discovery of the full extent of long and sparse structures, in particular, those of tidal nature. The difference between our selection of probable members and the datasets provided by two independent groups (Fürnkranz *et al.* 2019; Tang *et al.* 2019) is similar in scope to the difference between the datasets obtained in these papers. Partly they are due to different methods of filtering of data.

In further analysis we select a refined dataset of probable members of both stellar group and independently obtain their ages (700 ± 70 Myr and 450 ± 100 Myr) and space velocities $((U,V,W) = (8.63 \pm 0.13, 6.63 \pm 0.20, 6.65 \pm 0.95)$ km/s for Coma Ber star cluster, and $(U,V,W) = (7.70 \pm 0.12, 3.27 \pm 0.45, 5.69 \pm 0.80)$ km/s for Group X). Our results are in very good agreement with those of previous investigators, which again proves the validity of obtained results, and simultaneously provides independent constraints for the parameters of stellar groups.

In further work we intend to also use HDBSCAN for the same purpose and compare its effectiveness with DBSCAN, investigate the avoidance of false negatives in more detail by applying the method to mock stellar clusters and streams positioned at larger distances, and apply this method to larger samples of data, with the eventual goal of an all-sky survey of clusters and/or stellar comoving groups.

Acknowledgment: We are grateful to the anonymous referee for valuable comments and recommendations that resulted in significant improvement of this paper. This work has made use of data from the European Space Agency (ESA) mission Gaia processed by the Gaia Data Processing and Analysis Consortium (DPAC). Funding for the DPAC has been provided by national institutions, in particular, the institutions participating in the Gaia Multilateral Agreement. The use of TOPCAT (Taylor 2005) is acknowledged. The study was partially supported by Russian Foundation for Basic Research (Project no. 20-52-12009).

Author contributions: All authors have accepted responsibility for the entire content of this manuscript and approved its submission.

Conflict of interest: The authors state no conflict of interest.

References

- Bailer-Jones CAL. 2015. Estimating Distances from Parallaxes. *Publ Astron Soc Pac.* 127(956):994.
- Bressan A, Marigo P, Girardi L, Salasnich B, Dal Cero C, Rubele S, et al. 2012. PARSEC: stellar tracks and isochrones with the PADova and TRIeste Stellar Evolution Code. *Mon Not R Astron Soc.* 427(1):127–145.
- Brown AGA, Vallenari A, Prusti T, de Bruijne JHJ, Mignard F, Drimmel R, et al. 2016. Gaia Data Release 1. Summary of the astrometric, photometric, and survey properties. *Astron Astrophys.* 595:A2.
- Brown AGA, Vallenari A, Prusti T, de Bruijne JHJ, Babusiaux C, Biermann M, et al. 2021. Gaia Early Data Release 3. Summary of the contents and survey properties. *Astron Astrophys* 649, A1.
- Campello RJGB, Moulavi D, Sander J. 2013. Density-based clustering based on hierarchical density estimates. In Pei J, Tseng VS, Cao L, Motoda H, Xu G, Editors. *Advances in Knowledge Discovery and Data Mining*. Berlin, Heidelberg: Springer. p 160–172.
- Cantat-Gaudin T, Jordi C, Vallenari A, Bragaglia A, Balaguer-Nuñez L, Soubiran C, et al. 2018. A Gaia DR2 view of the open cluster population in the Milky Way. *Astron Astrophys.* 618:A93.
- Castro-Ginard A, Jordi C, Luri X, Álvarez Cid-Fuentes J, Casamiquela L, Anders F, et al. 2020. Hunting for open clusters in Gaia DR2: 582 new open clusters in the Galactic disc. *Astron Astrophys.* 635:A45.
- Castro-Ginard A, Jordi C, Luri X, Julbe F, Morvan M, Balaguer-Nuñez L, et al. 2018. A new method for unveiling open clusters in Gaia. New nearby open clusters confirmed by DR2 *Astron Astrophys.* 618:A59.
- El-Badry K, Rix H-W, Tian H, Ducheñe G, Moe M. 2019. Discovery of an equal-mass ‘twin’ binary population reaching 1000 + au separations. *Mon Not R Astron Soc.* 489(4):5822–5857.
- Ester M, Kriegel H-P, Sander J, Xu X. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In: *Proceedings of the Second International Conference*

- on Knowledge Discovery and Data Mining, KDD'96. AAAI Press, 1996. p. 226-231.
- Fuñkrantz V, Meingast S, Alves J. 2019. Extended stellar systems in the solar neighborhood. III. Like ships in the night: the Coma Berenices neighbor moving group. *Astron Astrophys.* 624:L11.
- Hunt EL, Reffert S. 2021. Improving the open cluster census. I. Comparison of clustering algorithms applied to Gaia DR2 data. *Astron Astrophys.* 646:A104.
- Jerabkova T, Bofl HMJ, Beccari G, de Marchi G, de Bruijne JHJ, Prusti T. 2021a. The 800 pc long tidal tails of the Hyades star cluster. Possible discovery of candidate epicyclic overdensities from an open star cluster. *Astron Astrophys.* 647:A137.
- Jerabkova T, Bofl HMJ, Beccari G, de Marchi G, de Bruijne JHJ, Prusti T. 2021b. The 800 pc long tidal tails of the Hyades star cluster. Possible discovery of candidate epicyclic overdensities from an open star cluster. *Astron Astrophys.* 647:A137.
- Jimeñez-Esteban FM, Solano E, Rodrigo C. 2019. A Catalog of Wide Binary and Multiple Systems of Bright Stars from Gaia-DR2 and the Virtual Observatory. *Astron J.* 157(2):78.
- Kamdar H, Conroy C, Ting Y-S, Bonaca A, Smith MC, Brown AGA. 2019. Stars that Move Together Were Born Together. *ApJ.* 884(2):L42.
- Kharchenko NV, Piskunov AE, Schilbach E, Rošer S, Scholz RD. 2012. Global survey of star clusters in the Milky Way. I. The pipeline and fundamental parameters in the second quadrant. *Astron Astrophys.* 543:A156.
- Meingast S, Alves J, Rottensteiner A. 2021. Extended stellar systems in the solar neighborhood. V. Discovery of coroneae of nearby star clusters. *Astron Astrophys.* 645:A84.
- Pang X, Li Y, Tang S-Y, Pasquato M, Kouwenhoven MBN. 2020. Different Fates of Young Star Clusters after Gas Expulsion. *ApJ.* 900(1):L4.
- Ratzenböck S, Meingast S, Alves J, Möller T, Bomze I. 2020. Extended stellar systems in the solar neighborhood. IV. Meingast 1: the most massive stellar stream in the solar neighborhood. *Astron Astrophys.* 639:A64.
- Rošer S, Schilbach E. 2019. Praesepe (NGC 2632) and its tidal tails. *Astron Astrophys.* 627:A4.
- Rošer S, Schilbach E. 2020. A census of the nearby Pisces-Eridanus stellar stream. Commonalities with and disparities from the Pleiades. *Astron Astrophys.* 638:A9.
- Rošer S, Schilbach E, Piskunov AE, Kharchenko NV, Scholz RD. 2011. A deep all-sky census of the Hyades. *Astron Astrophys.* 531:A92.
- Sapozhnikov SA, Kovaleva DA. 2020. Observational signs of wide binary stars dissolution in Gaia DR2 data. *INASANS Science Reports.* 5:336–340.
- Sapozhnikov SA, Kovaleva DA, Malkov OY, Sytov AY. 2020. Binary Star Population with Common Proper Motion in Gaia DR2. *Astron Rep.* 64(9):756–768.
- Schořrich R, Binney J, Dehnen W. 2010. Local kinematics and the local standard of rest. *MNRAS.* 403(4):1829.
- Tang S-Y, Pang X, Yuan Z, Chen WP, Hong J, Goldman B, et al. 2019. Discovery of Tidal Tails in Disrupting Open Clusters: Coma Berenices and a Neighbor Stellar Group. *ApJ.* 877(1):12.
- Taylor MB. 2005. TOPCAT & STIL: Starlink Table/VOTable Processing Software. In: Shopbell P, Britton M, Ebert R. Editors. *Proceedings of the Astronomical Data Analysis Software and Systems XIV ASP Conference Series.* 2004 Oct 24-27; Pasadena (CA), USA. ASPC, 2005. p.29.
- Yuan Z, Chang J, Banerjee P, Han J, Kang X, Smith MC. 2018. StarGO: A New Method to Identify the Galactic Origins of Halo Stars. *ApJ.* 863(1):26.
- Zhong J, Chen L, Kouwenhoven MBN, Li L, Shao Z, Hou J. 2019. Substructure and halo population of Double Cluster η and χ Persei. *Astron Astrophys.* 624:A34.