



Vera C. Rubin Observatory
Data Management

Host Galaxy Association for DIAObjects

M. L. Graham et al.

DMTN-151

Latest Revision: 2021-03-09



Abstract

This document describes the DIAObject parameters that associate extragalactic transients with potential host galaxies, and which will be generated during Alert Production and included in the alert packet.

Change Record

Version	Date	Description	Owner name
0	2020-02-26	Inception.	Melissa Graham
1	2020-05-11	Pre-RFC.	Melissa Graham
1.1	2021-03-09	Revision for DM-23683, RFC-695.	Melissa Graham

Document source location: <https://github.com/lstt-dm/dmtn-151>

Contents

1 Introduction	1
2 DIAObject–Object Association Parameters	1
2.1 Parameters for all transients and variables	1
2.2 Parameters for extragalactic transients	2
2.2.1 Calculating the separation distance.	2
2.2.2 Why these parameters are inadequate for low-redshift transients.	3
2.3 Parameters for low-redshift extragalactic transients	3
3 Options for Host Galaxy Association	4
3.1 Effective Radius	4
3.2 Second Moments	5
3.3 Hierarchical Associations	5
3.4 2D Algorithms and Footprints	6
3.5 Use of Galaxy/Transient Types	7
3.6 Identifying Hostless Transients	7
4 Appendix A: Separation Distance from Second Moments	8
5 Appendix B: Overcoming Background Galaxy Contamination	10
A References	14
B Acronyms	15

Host Galaxy Association for DIAObjects

1 Introduction

The LSST Alert Production pipeline will generate alert packets within 60 seconds of image readout for all sources detected during difference image analysis (DIA; DIASources). DIASources are associated by sky coordinate into DIAObjects, and a DIASource at a sky coordinate with no previous detections from DIA becomes a new DIAObject. The alert packet contains the entire DIASource and DIAObject record, including all the table elements described in the Data Products Definitions Document, LSE-163.

Individuals and brokers who process alerts will use the information in the packets to rapidly evaluate and prioritize DIAObjects for follow-up with limited resources. Thus, the contents of the alert packet need to contain a sufficient amount of LSST data about each DIAObject to enable immediate evaluation.

One important piece of information is the association of each DIAObject with a static-sky Object from the Data Release (DR) catalogs. Brokers will use the Object association to obtain data about the static-sky object from the DR catalogs, such as whether it might be galactic or extragalactic, at high- or low-redshift, nuclear or offset from a host, etc. Delivering alerts *with* the static-sky association already completed avoids the situation of multiple users cross-matching in real time.

This document describes the elements of the DIAObject catalog that contain the Object association parameters and describe how they will be calculated.

2 DIAObject-Object Association Parameters

2.1 Parameters for all transients and variables

The general parameters regarding associations between DIAObjects and Objects are described in LSE-163 as:

- `nearbyObj (unit64[6])`, the *"closest Objects (3 stars and 3 galaxies) in Data Release database"*

- nearbyObjDist (float[6]), the “distances to nearbyObj” in arcseconds
- nearbyObjLnP (float[6]), the “natural log of the probability that the observed DIAObject is the same as the nearby Object”

For nearbyObjLnP, a footnote states “This quantity will be computed by marginalizing over the product of position and proper motion error ellipses of the Object and DIAObject, assuming an appropriate prior”.

The definition of nearbyObjLnP is only appropriate for static variable point sources (stars): for transients in host galaxies, the observed DIAObject will never be “the same as the nearby Object”.

2.2 Parameters for extragalactic transients

There are a wide variety of options to associate extragalactic transients with their host galaxies, as presented in Section 3. Typically, radial distances in arcseconds such as nearbyObjDist, above, are not used in favor of a *separation distance* which accounts for the galaxies’ spatial luminosity profiles (e.g., Sullivan et al., 2006).

To associate DIAObjects with DR Objects, a separation distance calculated from the galaxies’ second moments (Section 3.2) will be used to generate DIAObject elements:

- potentialHost (unit64[3]), the objectId for the three Object catalog extended sources with the lowest separation distances
- potentialHostSeparation (float[3]), the separation distances for each potentialHost

2.2.1 Calculating the separation distance.

All of the parameters that are required in order to calculate the separation distances, as described in Appendix 4, will already exist in the DIAObject and Object tables. Separation distances need not be calculated for *all* Objects; a suitable limit would be to calculate separation distances for Objects within $\sim 90''$. A galaxy with a diameter of 25 kpc (i.e., like the Milky Way) has an angular diameter of

$\sim 30''$ at a distance of 200 Mpc ($z \sim 0.045$). Thus, considering all galaxies within $\sim 90''$ could associate a low-redshift transient at ~ 3 effective radii (R_e) from its host galaxy. There would be ~ 400 background galaxies within this $\sim 90''$ radius (scaled from the ~ 4 billion galaxies expected in the full $\sim 18000 \text{ deg}^2$ of the wide-fast-deep main survey), This means that just ~ 400 separation distances would need to be calculated for each DIAObject in order to fill the above table elements.

2.2.2 Why these parameters are inadequate for low-redshift transients.

In Appendix 5 we show that up to 1% of such low-redshift transients would randomly fall within $3R_e$ of three background interloper galaxies, and the true host would not be included in `potentialHost`. This issue is exacerbated for low-redshift transients further in the outskirts of their large host galaxies, or transients in the Local Volume – both of which have a very high scientific value for follow-up programs. Furthermore, large, extended, low-redshift galaxies are likely to be significantly fragmented in the DR Objects catalog, with plenty of “child” Objects that would complicate host association. These issues with low-redshift host association *could* be mitigated by reporting a larger number of nearby potential hosts (e.g., Section 3.3), or by dedicating significant (and currently out of scope) effort to building a low-redshift DR Object catalog. However, Section 2.3 presents a more straightforward solution to low-redshift host association.

2.3 Parameters for low-redshift extragalactic transients

An established catalog of low-redshift galaxies (e.g., from the NGC/IC list of $\lesssim 14000$ Revised New General Catalogue and Index Catalogue; TBD) will be adopted, and all new DIAObjects will be cross-matched to it during Alert Production in order to report the name and the radial sky separation of the nearest low-redshift galaxy:

- `potentialLowzHost (str[1])`, the catalog name of the nearest ?low-z? potential host
- `potentialLowzHostSeparation (float[1])`, the distance to `potentialLowzHost` [arcsec]

If the community provides a catalog of names and coordinates that they would prefer to be used, it will be adopted. If that catalog includes more information about the galaxy, then the value of `potentialLowzHostSeparation` could be more sophisticated than the radial distance in

arcseconds. For example, if distances are included then the offset in kpc could be provided; if sizes are included the number of effective radii could be reported; or if the second moments are included then the separation could be the same as used for `potentialHostSeparation`.

3 Options for Host Galaxy Association

Statistically, the most likely host for a given transient is the galaxy which contributes the most optical flux at the transient's location. This is usually estimated by first calculating a *separation distance* from the nearby galaxies to the transient which is expressed in terms of the galaxy's spatial luminosity profile, and then assuming the galaxy with the lowest separation distance is the host.

However, there are many different ways to evaluate a separation distance, such as using the galaxy's effective radius (Section 3.1) or second moments (3.2) – potentially on multiple size scales (??) – or using extended source footprints (3.4).

All of these options are discussed below as context for the alert packet parameters presented in Section 2.

Additionally, advanced techniques that include the type of galaxy and transient as priors in probabilistic host galaxy association are briefly described in Section 3.5, and criteria for declaring a transient to be hostless is briefly discussed in Section 3.6. Neither of these types of evaluations will be incorporated into the `DIAObject` parameters because they are more sensitive to the particular use-cases, and so are left for scientists to implement.

3.1 Effective Radius

The effective (or half-light) radius of a galaxy is a standard measure of its size. The separation distance can be calculated by taking the radial distance from the core of the galaxy to the location of the transient, and dividing it by the effective radius of the galaxy. In this case the separation distance can be compared between potential hosts because it has been normalized by the potential hosts' sizes. The `DR Objects` table is already planned to contain suitable effective radii such as the parameter `kronRad90` [LSE-163], so it is a feasible option. This option would not require any additional processing aside from dividing the radial distance from transient to `Object` by the effective radius of the `Object`.

Although this kind of separation distance would account for the relative sizes of the potential host galaxies, it does not account for their position angles, and so would not be as accurate for assessing potential high-inclination host galaxies as using the second moments, as described in Section 3.2.

3.2 Second Moments

This method calculates a separation distance based on the two-dimensional luminosity profile of the nearby galaxies. For example, Sullivan et al. (2006) describe the method applied to the Supernova Legacy Survey (SNLS), using a separation distance of $R^2 = C_{xx}x_r^2 + C_{yy}y_r^2 + C_{xy}x_ry_r$, where C_{xx} , C_{yy} , and C_{xy} are ellipse parameters derived from the second moments of the galaxy luminosity profile and x_r, y_r are the on-sky distances between the centroids of the transient and the galaxy. The DR Objects table is already planned to contain the second moments of the galaxy luminosity profiles (I_{xx} , I_{yy} , and I_{xy} ; LSE-163). A step-by-step description of how this separation distance can be calculated from planned DIAObject and Object table elements is provided in Section 4.

Multiple recent surveys have used this method (or similar) to associate transients with their host galaxies, such as Sako et al. (2018) for SDSS supernovae, and Gupta et al. (2016) who use real and simulated data to evaluate the optimal method for host association. This option for the separation distance requires more computational steps than effective radius, but accounts for both the relative sizes and position angles of the potential host galaxies.

3.3 Hierarchical Associations

Instead of associating a DIAObject with the three extended Objects with the lowest separation distance, another option is to associate the DIAObject with the nearest Object at three different size scales. For example, the nearest $R_e < 10''$ neighbor within $d < 100''$ (high- z and small galaxies), the nearest $R_e < 100''$ neighbor within $d < 1000''$ (large low- z galaxies), and the nearest $R_e < 1000''$ neighbor within $d < 10000''$ (very large nearby galaxies). In this option, the “nearest neighbor” in a given scale bin could be determined using a separation distance calculated from, e.g., the second moments (Section 3.2).

However, this option does not avoid the issue of contamination by background galaxies as discussed in Appendix 5. To mitigate background interlopers, the nearest *three* extended

sources for *each* size scale should be included. This is less necessary for the largest size scale, but it would assist with identifying transients in galaxy groups and clusters, especially the rare transients with large offsets which might belong to intracluster stellar populations.

This option of hierarchical associations would require three times as much storage space in the DIAObject catalog. Although this option might reduce the number of Objects for which a separation distance must be calculated for each DIAObject, as described in Section 2.2 this number is only ~ 400 (i.e., not overwhelmingly large).

Ultimately, the largest size scale would face the same issues for low-redshift galaxies as described in Section 2.2.2. This method of hierarchical associations is thus not adopted for the Alert Production pipeline, as the combination of the methods described in Section 2.2 and 2.3 will, together, adequately cover the science needs.

3.4 2D Algorithms and Footprints

Aside from adopting a separation distance, there are more complicated methods for identifying the most likely host for a given transient. For example, the nearby galaxy with the smallest fraction of light interior to an isophot through the transient's location, where the isophot shape is given more degrees of freedom and not constrained to concentric ellipses as in the second moment method above. Another example is to use an algorithm that provides deblended footprints for nearby extended objects, and that can estimate the fraction of light in given pixel that should be attributed to each (e.g., as the SCARLET deblender can do, Melchior et al. (2018)). The most likely host galaxy would be the one which contributes the most flux at the pixel location of a transient. While all DR Objects will be associated with a footprint (a region of connected pixels), the footprint information will not be stored in the Object table. The use of footprints in identifying potential host galaxies would require more computational resources during Alert Production, but would probably return a more accurate host association for only a very small fraction of DIAObjects.

Methods based on 2D algorithms or footprints are thus not adopted for the Alert Production pipeline, but might be applied by science users in their analysis of the Data Release data products.

3.5 Use of Galaxy/Transient Types

The association of transients with their host galaxy can be more accurate if their properties are also considered. For example, the potential host galaxy's redshift can be used to calculate separation distances in physical units, or to estimate the absolute brightness of the transient and consider whether it is physically plausible. Priors based on the established correlations between transient types and host galaxy morphology or color can also be used to refine a probabilistic host association, such as how core collapse supernovae are almost always associated with star formation (except for a few notable cases, e.g., Graham et al. 2012; Irani et al. 2019). A demonstration that these correlations between host and transient types are so robust that the host type can be used to provide a statistical classification of the transient type was presented by Foley & Mandel (2013).

However, making science-informed associations between transients and galaxies based on any kind of derived properties is beyond the scope of Alert Production, and is best left to the users on a case-by-case basis. Thus, properties of the transients and/or the nearby galaxies (beyond their coordinates and luminosity profile) should not be used during the association of DIAObjects and Objects.

For a very thorough assessment of an optimized, science-driven system for associating supernovae and their host galaxies – including the role and performance of machine learning methods – we direct the reader to Gupta et al. (2016).

3.6 Identifying Hostless Transients

For some scientific analyses, transients which are $>3-5$ effective radii away from the nearest galaxy, or for which $> 99\%$ of the potential host's luminosity is within the radial distance between transient and host center, are considered "hostless" (e.g., Sand et al. 2011). Such a cutoff has been appropriate for past samples of \sim hundreds of transients, but will not be appropriate for the LSST sample size. Furthermore, the decision of whether and how to consider a transient "hostless" is best left as a scientific decision for the end-user. Thus, no such cut should be applied during the association of DIAObjects and Objects, and the most likely hosts should still be reported, even if the probability is low.

4 Appendix A: Separation Distance from Second Moments

From the LSST catalogs the following Object table elements are used to define the parameters needed to calculate the separation distance [LSE-163]:

Parameter	Unit	Table Element	Description
$x_{\text{trans}}, y_{\text{trans}}$	degrees	DIAObject radec	transient centroid
$\overline{x}_{\text{gal}}, \overline{y}_{\text{gal}}$	degrees	Object radec	galaxy centroid
$\overline{x^2}, \overline{y^2}, \overline{xy}$	arcsec ²	Object Ixx, Iyy, Ixy	galaxy second moments

There might be an issue with using the Object catalog second moments: the Ixx, Iyy, and Ixy are defined with respect to the local tract/patch and not sky coordinates (Jira DM-19519). Although the local tangent projection will probably work fine for this application of the second moments, this should be verified at the time of implementation.

As described in Section 10 of E. Bertin's Source Extractor manual¹ (and presumably many other places), the unitless ellipse parameters C_{xx} , C_{yy} , C_{xy} can be calculated from the second moments via:

$$C_{xx} = \frac{\overline{y^2}}{\sqrt{\left(\frac{\overline{x^2} - \overline{y^2}}{2}\right)^2 + \overline{xy}^2}} \quad (1)$$

$$C_{yy} = \frac{\overline{x^2}}{\sqrt{\left(\frac{\overline{x^2} - \overline{y^2}}{2}\right)^2 + \overline{xy}^2}} \quad (2)$$

$$C_{xy} = -2 \frac{\overline{xy}}{\sqrt{\left(\frac{\overline{x^2} - \overline{y^2}}{2}\right)^2 + \overline{xy}^2}} \quad (3)$$

The sky distances between the transient and galaxy centroids are calculated as follows, and

¹Version 2.3: <https://www.astromatic.net/pubsvn/software/sextractor/trunk/doc/sextractor.pdf>

include the $\cos(\delta)$ factor and a conversion from units of degrees to arcseconds:

$$x_r = 3600(x_{\text{SN}} - x_{\text{gal}}) \quad (4)$$

$$y_r = 3600(y_{\text{SN}} - y_{\text{gal}}) \cos y_{\text{gal}} \quad (5)$$

Finally, the separation distance R in arcseconds is given by:

$$R^2 = C_{xx}x_r^2 + C_{yy}y_r^2 + C_{xy}x_r y_r. \quad (6)$$

5 Appendix B: Overcoming Background Galaxy Contamination

In this appendix, the probability of failed host galaxy associations for nearby transients with large host offsets due to interloping background galaxies is quantified, and used to evaluate the minimum number of potential hosts that should be stored in a DIAObject record. For simplicity, this appendix uses the effective radius as the separation distance (Section 3.1).

The 10-year Object catalogs will include ~ 4 billion galaxies with $i < 25$ mag across the ~ 20000 deg² main survey area, known as the “gold” sample, especially in the context of weak lensing studies [LPM-17]. However, the 10-year coadded depths will detect galaxies down to 5σ limiting magnitudes of 26.1, 27.4, 27.5, 26.8, 26.1, and 24.9 mag in filters *ugrizy*; this is ~ 3 times as many as in the “gold” sample, or ~ 10 billion galaxies. This high density of background galaxies complicates the process for associating large nearby host galaxies with their transients, especially the rare transients in their outskirts.

Consider a transient at $3R_e$ from the center of a nearby galaxy with $z = 0.01$ and an effective radius of $R_e = 10$ kpc (~ 50 arcsec). In order for this transient to be associated with its true host, the separation distance for all Objects within a radius of at least $3R_e$, and thus an area of at least $A_{3R_e} = \pi(3R_e)^2 = 0.0052$ deg², would need to be considered. Based on the final 10-year number of detected galaxies (10 billion) and the total survey area (20000 deg²), that is $\sim (10^{10}/20000) \times 0.0052 \approx 2600$ galaxies. Furthermore, the true host galaxy must have a lower separation distance than the N nearest background galaxies, where N is the number of potential hosts that will be listed in the DIAObject parameters `potentialHost` and `potentialHostSeparation`.

To investigate the probability of host-association failure for nearby transients, we simulate a mock catalog of randomly distributed random background interloper galaxies. It is based on the same LSST-like mock galaxy catalog used by Graham et al. (2018) for studies of photometric redshifts. For this experiment the catalog is limited to galaxies with at least a 5σ detection in the *griz* bands at the projected 10-year depths listed above. The catalog contains redshifts and apparent *ugrizy* magnitudes, which are used to approximate intrinsic absolute *r*-band magnitudes (without *K*-correction, simply using a distance modulus based on redshift and $M_r = m_r - \mu$). The absolute magnitudes are used to synthesize approximate galaxy radii based on the relationship between absolute magnitude and radius for late-type galaxies defined in Figure 3 of Shen et al. (2003). These synthesized radii are over-estimates because late-types are generally larger than early-types, and because this magnitude-radius relation

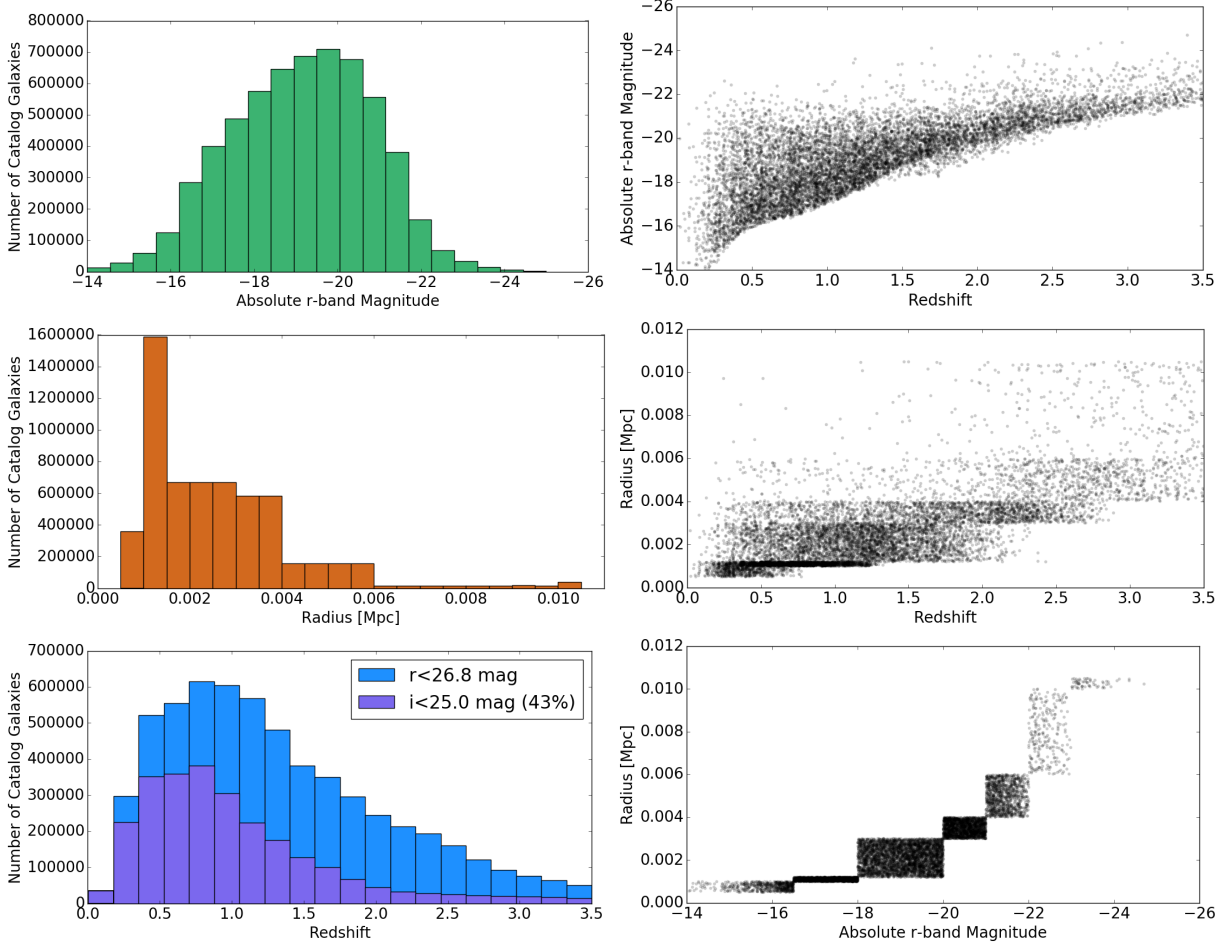


FIGURE 1: *Left:* Histograms of the synthesized absolute r -band intrinsic magnitude (top) and radius (middle) for catalog galaxies, and the simulated galaxy redshifts (bottom). *Right:* Correlation with redshift of the synthesized intrinsic magnitude (top) and radius (middle), and the approximate relation between radius and intrinsic magnitude (bottom).

was defined for SDSS galaxies at lower redshifts than the LSST high- z galaxies it is being applied to. This is a deliberate choice to overestimate, because it will result in upper limits on the rate of interlopers, which will allow for conservative estimates about the effect of background interlopers. The characteristics of this crude galaxy catalog are illustrated in Figure 1.

Consider again the large nearby galaxy with $R_e = 10$ kpc at a redshift of $z = 0.01$, for which the sky area within $3R_e$ is $A(3R_e) = 0.0052 \text{ deg}^2$ and contains ~ 2600 background galaxies. From the simulated catalog described above, 100 sets of background galaxies are randomly selected. For each set, the fraction of the nearby galaxy's area which is covered by the area of interloping background galaxies is calculated: $f_A = \sum A(3R_{e,\text{bkg}})/0.0052$. This fraction is equivalent to the

probability that a transient at $3R_e$ from this large nearby host will be within $3R_{e,\text{bkg}}$ (i.e., will be closer to) a background interloper. The probability that this transient is closer to N interlopers than to its true host is $P_{\text{fail}} = (f_A)^N$. This is the probability of a failed host association, where “failed” means that the DIAObject record of the N galaxies with the lowest separation distances does not include the true host. For this large nearby galaxy, average values of f_A and P_{fail} from the 100 simulated background sets are $f_A = 0.181 \pm 0.008$ and $P_{\text{fail}}(N = 3) = 0.006 \pm 0.0008$

Since this probability of failure is based on the sky density of background galaxies, it is independent of the radius and redshift of the nearby galaxy. However, it does depend on the factor applied to effective radius (i.e., the offset of transient considered) and N , the number of potential hosts recorded in the DIAObject catalog. Figure 2 shows the probability of failure as a function of R_e and N . **These are upper limits on the probability of failure** because they are derived from upper-estimates of galaxy radii and the sky density of the final 10-year LSST galaxy catalog, as described above.

To assist in the interpretation of Figure 2, the following describes several conclusions drawn from points in this plot:

- Transients with low host offsets, $1R_e$, are closer to (within $1R_{e,\text{bkg}}$ of) one background interloper $\sim 2\%$ of the time (purple point at $1R_e$). (However, given the high surface brightness of nearby galaxies, such a background galaxy might not be detected.)
- Transients with high host offsets, $5R_e$, are closer to (within $5R_{e,\text{bkg}}$ of) one background interloper $\sim 50\%$ of the time (purple point at $5R_e$), and closer to six background interlopers $> 1\%$ of the time (magenta point at $5R_e$).
- In order to achieve a 1% probability of failure for transients offset by $3R_e$ from large nearby galaxies, the DIAObject catalog record should include the $N = 3$ galaxies with the lowest separation distances (green point at $3R_e$).

The above experiment could be summarized in two main take-away points for reducing the probability of failure in associating nearby, large-offset transients with their true hosts:

1. The separation distances for all galaxies within at least $4R_e \approx 200$ arcsec should be calculated and considered.

2. The 3 galaxies with the lowest separation distances should be included in the DIAObject catalog record.

Adopting these recommendations would cause up to 1% of the transients at $3R_e$ from large nearby galaxies to experience a failed host association, where the true host is not listed in the DIAObject record. Since $3R_e$ encompasses $\sim 99\%$ of a galaxy's light, and most transient types are distributed proportional to the light, **the upper limit on the host association failure rate** for nearby transients should be $\sim 0.01\%$.

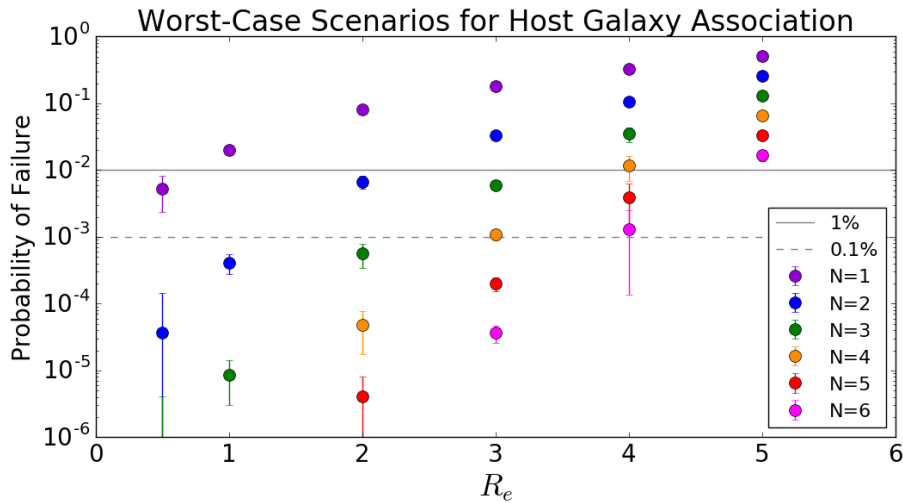


FIGURE 2: The probability that a transient will fail to be associated with a large ($R_e = 10$ kpc) nearby host galaxy due to background interlopers, as a function of the transient's offset in effective radii from galaxies in the vicinity (including the true host), where "failure" means the true host's separation distance is not in the top N nearest galaxies. **This is a "worst case scenario" because it applies to an very large nearby galaxy, and all background galaxy radii estimates are upper limits** (as described in the text). Error bars show the standard deviation from the 100 randomly-generated sets of background galaxies.

A References

References

- Foley, R.J., Mandel, K., 2013, *ApJ*, 778, 167 (arXiv:1309.2630), doi:10.1088/0004-637X/778/2/167, ADS Link
- Graham, M.L., Sand, D.J., Bildfell, C.J., et al., 2012, *ApJ*, 753, 68 (arXiv:1205.0015), doi:10.1088/0004-637X/753/1/68, ADS Link
- Graham, M.L., Connolly, A.J., Ivezić, Ž., et al., 2018, *AJ*, 155, 1 (arXiv:1706.09507), doi:10.3847/1538-3881/aa99d4, ADS Link
- Gupta, R.R., Kuhlmann, S., Kovacs, E., et al., 2016, *AJ*, 152, 154 (arXiv:1604.06138), doi:10.3847/0004-6256/152/6/154, ADS Link
- Irani, I., Schulze, S., Gal-Yam, A., et al., 2019, *ApJ*, 887, 127 (arXiv:1904.01425), doi:10.3847/1538-4357/ab505d, ADS Link
- [LPM-17]**, Ivezić, Ž., The LSST Science Collaboration, 2018, *LSST Science Requirements Document*, LPM-17, URL <https://ls.st/LPM-17>
- [LSE-163]**, Jurić, M., et al., 2017, *LSST Data Products Definition Document*, LSE-163, URL <https://ls.st/LSE-163>
- Melchior, P., Moolekamp, F., Jerdee, M., et al., 2018, *Astronomy and Computing*, 24, 129 (arXiv:1802.10157), doi:10.1016/j.ascom.2018.07.001, ADS Link
- Sako, M., Bassett, B., Becker, A.C., et al., 2018, *PASP*, 130, 064002 (arXiv:1401.3317), doi:10.1088/1538-3873/aab4e0, ADS Link
- Sand, D.J., Graham, M.L., Bildfell, C., et al., 2011, *ApJ*, 729, 142 (arXiv:1011.1310), doi:10.1088/0004-637X/729/2/142, ADS Link
- Shen, S., Mo, H.J., White, S.D.M., et al., 2003, *MNRAS*, 343, 978 (arXiv:astro-ph/0301527), doi:10.1046/j.1365-8711.2003.06740.x, ADS Link
- Sullivan, M., Le Borgne, D., Pritchet, C.J., et al., 2006, *ApJ*, 648, 868 (arXiv:astro-ph/0605455), doi:10.1086/506137, ADS Link

B Acronyms