PROCEEDINGS OF SCIENCE

A Graph-based Hierarchical Clustering Algorithm for population studies of astrophysical objects

Jonathan Mauro,[a,*] **Karlijn Kruiswijk,**[a] **Emile Moyaux,**[a] **Christoph Raab**[a] **and Gwenhaël W. de Wasseige**[a]

[a] *Centre for Cosmology, Particle Physics and Phenomenology - CP3,*
Universite Catholique de Louvain, B-1348 Louvain-la-Neuve, Belgium

E-mail: jonathan.mauro@uclouvain.be

We present here a data exploration tool designed to enhance the study of astrophysical objects by integrating traditional hierarchical clustering with graph-based community detection algorithms. This new tool allows in-depth analysis of the distributions of observables across astrophysical catalogs. The method is first validated on the Iris benchmark dataset, where it accurately reproduces the known taxonomic classification and outperforms default implementations of several widely used clustering algorithms. We then showcase applications to the SWIFT catalog of gamma-ray bursts. Finally, we discuss the most representative features that characterise the identified subpopulations to allow for qualitative analysis and model development.

39th International Cosmic Ray Conference (ICRC2025)
15–24 July 2025
Geneva, Switzerland

ICRC 2025
The Astroparticle Physics Conference
Geneva July 15-24, 2025

[*] Speaker

https://pos.sissa.it/

1. Motivation

In neutrino astronomy, stacking is a widely used technique that increases statistical power by combining signal from a population of sources. This approach relies on the assumption that all sources within a given catalog share a common underlying physical process that can be probed collectively. Although this assumption is often not expected to be valid, stacking is generally regarded as a sensible alternative to source-by-source analyses, especially due to a limited understanding of the potential emission mechanism.

In these proceedings we propose a third approach: subdividing the catalog into smaller groups based on similarities in observational data, and performing stacking within each group. This strategy aims to relax the strong assumption of uniformity across the entire catalog by replacing it with a more reasonable hypothesis—that sources with similar experimental signatures are likely to share the same production mechanisms.

The advantage of this approach is that it enables more reliable constraints when comparing with emission models that correlate with observational signatures. For example, if a neutrino emission model predicts stronger signals from sources embedded in dense environments, it is more appropriate to compare the model with limits derived from gamma-ray–dim sources, which are more likely to represent the relevant population.

With this in mind, we introduce a novel clustering algorithm designed to identify such subpopulations with astrophysical catalogs. The Graph-based Hierarchical Clustering Algorithm (GHCA) provides a framework for applying graph community detection techniques to data that are not inherently graph-structured. We begin by presenting an overview of the algorithm and then demonstrate its performance on the iris benchmark dataset. Finally, we apply GHCA to an astrophysical catalog.

2. Graph-based Hierarchical Clustering Algorithm

GHCA provides an interface between the community detection algorithms available in NetworkX [1] and scikit-learn clustering API [2, 3]. To enable this integration, the input data must be encoded as a graph, where each data point corresponds to a node. Unlike other approaches that rely on k-nearest neighbor graphs, GHCA constructs distance-based edges: two nodes are connected if the pairwise distance between the corresponding data points is smaller than a threshold d_t, as illustrated in Fig. 1. Any pairwise distance metric can be used; by default GHCA uses a cosine distance defined as $d_{\cos}(\mathbf{X}, \mathbf{Y}) = 1 - \frac{\langle \mathbf{X}, \mathbf{Y} \rangle}{\|\mathbf{X}\| \|\mathbf{Y}\|}$, where $\mathbf{X}$ and $\mathbf{Y}$ are two data points and $\langle \mathbf{X}, \mathbf{Y} \rangle$ their dot product, but other metrics are supported, e.g., euclidean distance or the so-called manhattan distance, given by the sum of element-wise absolute differences.

Clusters are identified using community detection algorithms implemented in NetworkX, such as the Louvain algorithm [4], which is the default choice in GHCA, or the Leiden algorithm [5]. In the simplest case, if the connected components of the graph are treated as communities—e.g., as presented in [6]—the method reproduces the behaviour of traditional single-linkage hierarchical clustering [7].

As with most hierarchical clustering algorithms, GHCA produces a partition for any given distance threshold. To explore the structure of the data at different levels of connectivity, the algorithm is typically evaluated multiple times using an increasing sequence of distance thresholds d_t, selected from the set of equally spaced quantiles of the distribution of all pairwise distances. This approach ensures a consistent progression of graph connectivity across evaluations. For instance, if only one evaluation is performed, the chosen threshold corresponds to the median of the pairwise distance distribution, resulting in half of the total pairs of nodes being connected. If two evaluations are performed instead, the thresholds are chosen such that the two resulting graphs will have one-third and two-thirds of the pairs connected. This procedure generalizes naturally for a higher number of evaluations, enabling systematic exploration of the clustering structure across different graph densities.

When a number of clusters is specified, GHCA returns the partition found by the community detection algorithm that matches such number if one exists, if no such partition is found, it is interpreted as an indication that the chosen number of cluster is not suitable for the dataset. In the majority of cases however, multiple distinct partitions yield the same number of clusters, when this occurs, GHCA selects the one with a higher silhouette score—a clustering quality metric that generally favors convex-shaped clusters. Accordingly, if the number of clusters is not specified, GHCA will return the partition with highest score independently of the number of clusters found.

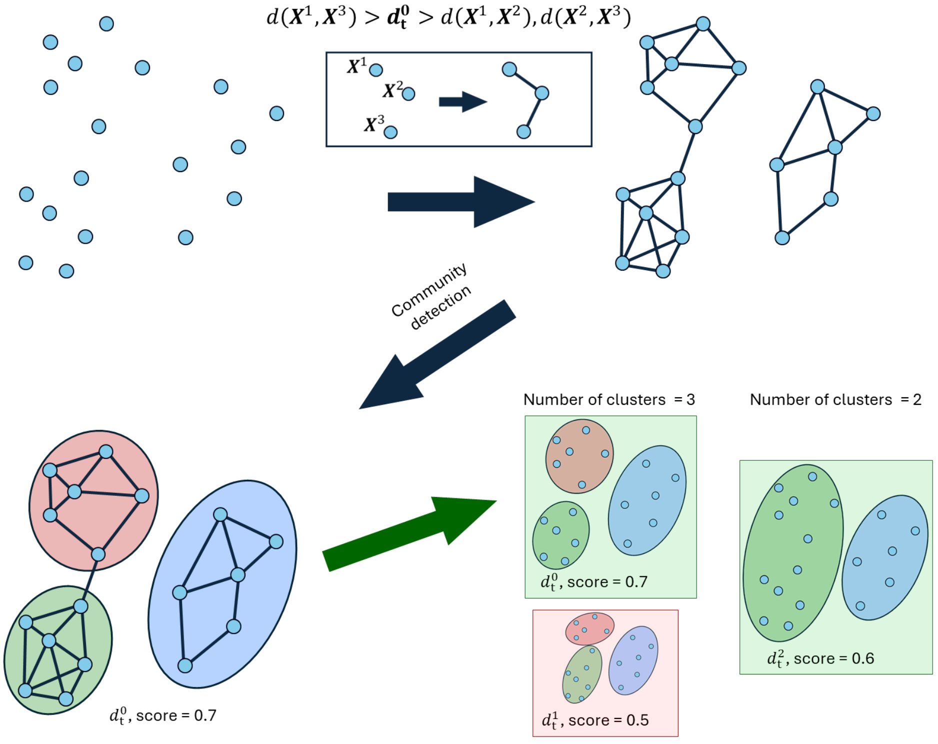
$d(\boldsymbol{X}^1,\boldsymbol{X}^3) > \boldsymbol{d}_\mathrm{t}^0 > d(\boldsymbol{X}^1,\boldsymbol{X}^2), d(\boldsymbol{X}^2,\boldsymbol{X}^3)$
$\boldsymbol{X}^1$
$\boldsymbol{X}^2$
$\boldsymbol{X}^3$
Community detection
Number of clusters = 3
Number of clusters = 2
d_t^0, score = 0.7
d_t^1, score = 0.5
d_t^2, score = 0.6
d_t^0, score = 0.7

Figure 1: Diagram illustrating the steps of the GHCA algorithm during the evaluation of a distance threshold value d_t^0. Following the flow indicated by the black arrows, the figure shows the criterion used to define graph edges and subsequently the identification of communities (red, blue and green ovals). For illustrative purposes, the diagram also compares the outcomes of multiple threshold evaluations—d_t^0, d_t^1, and d_t^2—and highlights that for a given number of identified clusters, the configuration with higher score is preferred (green shaded background) over the one with lower score (red shaded background). The communities and scores are hypothetical.

3. Iris dataset

To evaluate the performance of the algorithm, benchmark datasets can be employed. In these proceedings we showcase one of the most widely used examples: the Iris dataset [8]. This dataset contains measurements of petal and sepal dimensions from three species of the Iris flower—Iris setosa, Iris virginica, and Iris versicolor— with 50 samples per species. Since the true species labels are known, the dataset serves as a valuable reference for assessing the effectiveness of unsupervised clustering algorithms by comparing the inferred partitions against the taxonomic classification.

Figure 2 shows GHCA results on the iris dataset projected on the parameter space of its first two principal components (PC). For reference, the ground truth labels are also shown, alongside the partitions produced by several widely used clustering algorithms: KMeans [9], HDBSCAN [10], average-linkage hierarchical clustering (agglomerative clustering), and a Gaussian Mixture Model [11]. All algorithms were executed using their default parameter settings from the scikit-learn implementation. However, the number of clusters was explicitly set to three in each case—except for HDBSCAN, which does not use this parameter, therefore, found a bipartition of the dataset. In this evaluation, GHCA was run over 30 distance thresholds.

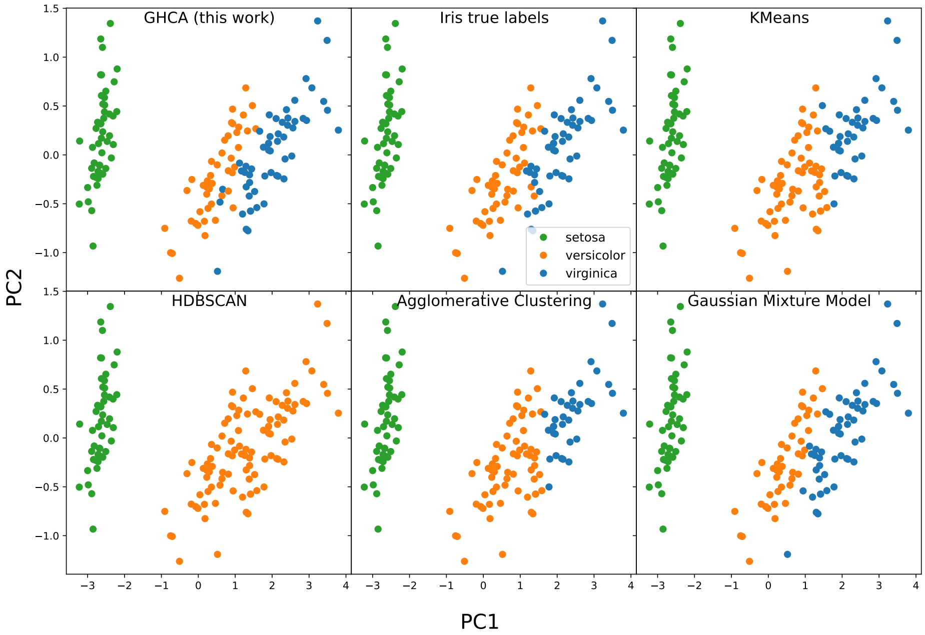

Figure 2: Comparison of clustering results for the Iris dataset. The top row (left to right) shows the results from GHCA, the ground truth species classification, and KMeans. The bottom row shows the results from HDBSCAN, Agglomerative Clustering, and Gaussian Mixture model. Default implementations of each algorithm were used, and where applicable, the number of cluster was set to three. Clusters are indicated by using different colours.

The three clusters identified by GHCA closely match the taxonomic classification and are comparable in accuracy to the predictions of the Gaussian Mixture Model. In contrast, KMeans and agglomerative clustering algorithms fail to correctly separate the versicolor and virginica species. The bipartition produced by HDBSCAN merges these two species, while isolating setosa as a separate group. This outcome reflects the greater similarity between versicolor and virginica as opposed to the setosa, which is more clearly separated in the parameter space.

Although the test described here provides insight into the different behaviors and performances of various clustering algorithms, it does not serve as an absolute comparison. The accuracy with which each algorithm identifies the underlying true classes can vary significantly depending on the shape and distribution of those classes within the parameter space used for clustering.

4. Gamma-ray bursts catalog

Here, we discuss the application of GHCA to the catalog of gamma-ray bursts (GRBs) observed by SWIFT [12, 13]. The data for this demonstration were obtained via the open-source module provided by the UK Swift Science Data Centre[1]. Since our aim is to highlight patterns reflecting different GRB progenitors or environments, we consider only measurements related to emission mechanisms, discarding positional and temporal information.

Observations were performed by the BAT or XRT [14] instruments onboard SWIFT, covering soft gamma rays (15–150 keV) and X-rays (0.3–10 keV), respectively. Specifically, we use the fitted spectral indices γ from both BAT and XRT, the BAT fluence and peak flux, the XRT flux measured over the first 11 hours after the burst, and the XRT initial temporal decay index α. Only GRBs with all these measurements were selected, yielding a sample of 700 bursts.

To provide a more meaningful notion of distance in parameter space, the data were preprocessed using scikit-learn's RobustScaler, which mitigates the influence of outliers. Additionally, the logarithms of the BAT fluence and the XRT 11-hour flux were used instead of their nominal values.

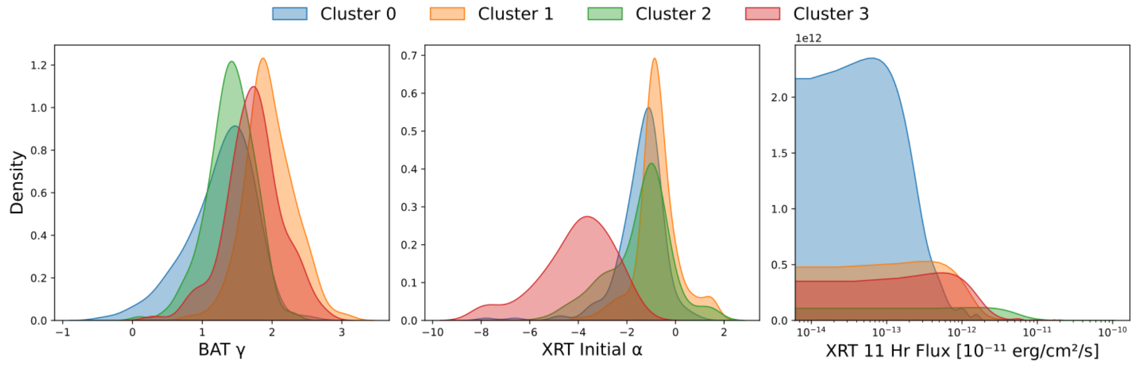

Figure 3: Density distributions of identified clusters of GRBs in the soft-gamma-ray spectral index: BAT γ, X-rays initial temporal decay index: XRT Initial α, and in the XRT 11-hour flux.

Clustering results for this GRB sample are shown in Fig. 3 and Fig. 4. Using the default GHCA implementation described in Section 2, four clusters were identified, each with comparable sizes: 196, 162, 203, and 193 GRBs. Characteristic observational features of each cluster can be readily identified from their per-cluster distributions. For example, Fig. 3 highlights three features—BAT spectral index, XRT initial temporal decay index, and XRT 11-hour flux—revealing that Cluster 3 consists of GRBs with low initial decay indices and high spectral indices and 11-hour fluxes.

[1] https://www.swift.ac.uk/API

Conversely, Cluster 0 GRBs exhibit low spectral index and 11-hour flux, combined with high initial decay index.

To further illustrate these relationships, Fig. 4 presents the clusters in two-dimensional projections of all measurement pairs. This visualization confirms that no strong correlations exist between the observables used. For population studies of this kind, it is recommended to avoid strongly correlated measurements, as they can bias the algorithm toward identifying clusters primarily in the space of correlated observables.

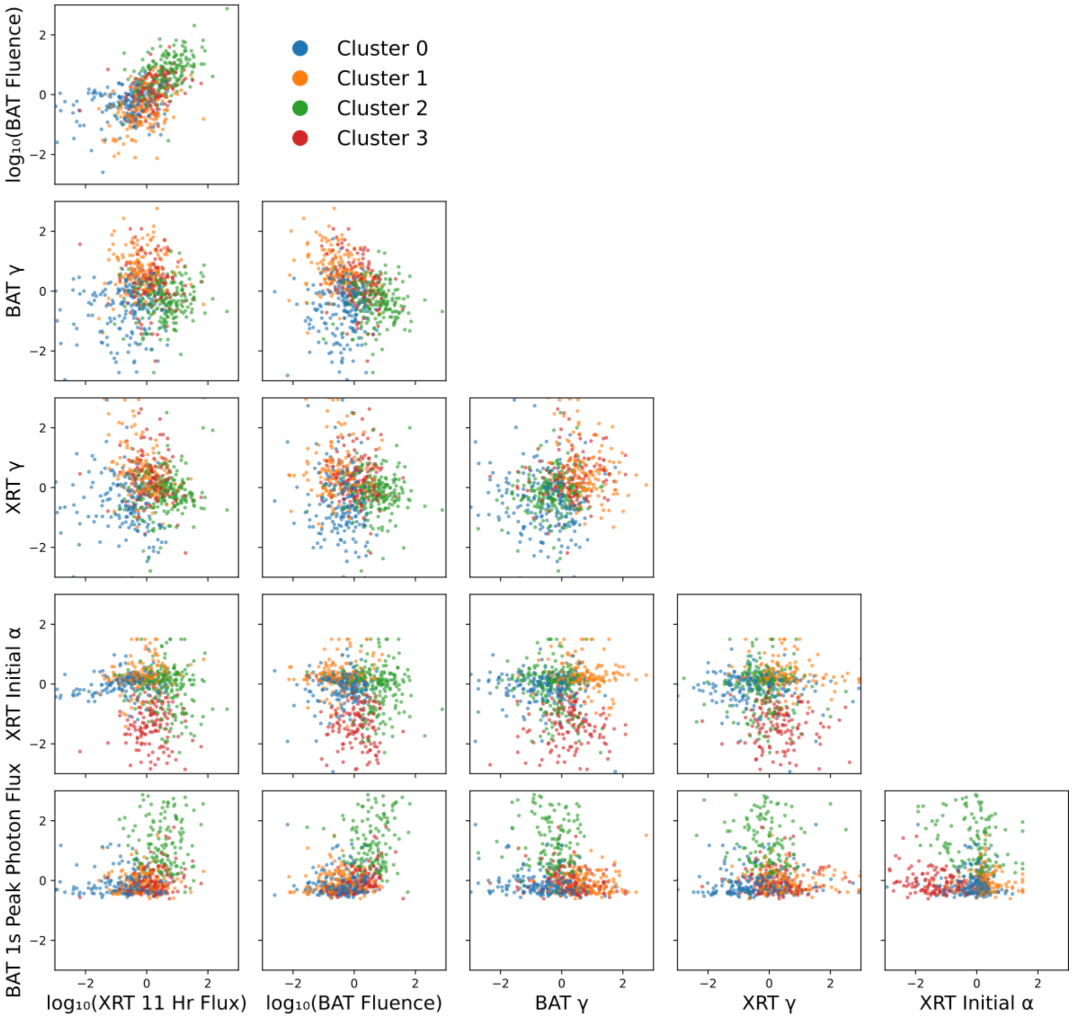

Figure 4: GRBs projected onto two-dimensional spaces for each pair of observables used in this analysis. Different colors indicate the clusters identified by GHCA. Observations were scaled using scikit-learn's RobustScaler; thus, zero corresponds to the median of each observable, and values are rescaled by the interquartile range.

5. Conclusions

These proceedings present an argument for employing clustering algorithms to define source populations in astrophysical catalogs, with the aim of enabling population-specific neutrino searches.

The algorithm introduced here—GHCA—leverages community detection methods developed for graph analysis and applies them to data that are not inherently graph-structured, such as catalogs of astrophysical sources. GHCA has been designed to be fully compatible with the scikit-learn API, ensuring ease of use and integration.

On the Iris benchmark dataset, GHCA demonstrates strong performance, comparable to that of a Gaussian Mixture Model, and achieves more accurate classifications than widely used algorithms such as KMeans, HDBSCAN, and agglomerative clustering. We then apply GHCA to identify meaningful populations within the SWIFT GRB catalog, using measurements from the BAT and XRT instruments in the soft gamma-ray and X-ray bands. More extensive studies are ongoing for the use of GHCA on astrophysical catalogs, in particular, Fermi GRB catalog, the Fermi-LAT Solar Flare catalog, compact binary mergers detected through gravitational waves and active galactic nuclei.

Acknowledgments

JM is a PhD fellow of FRIA du Fonds de la Recherche Scientifique - FNRS, KK and EM are Research Fellows of the Fonds de la Recherche Scientifique - FNRS. GDW acknowledges support by the Francqui Foundation. This work was supported by the Fonds de la Recherche Scientifique - FNRS under Grant no. 4.4501.02. This work made use of data supplied by the UK Swift Science Data Centre at the University of Leicester.

References

[1] A.A. Hagberg, D.A. Schult and P.J. Swart, *Exploring network structure, dynamics, and function using networkx*, in *Proceedings of the 7th Python in Science Conference*, (Pasadena, CA USA), pp. 11 – 15, 2008.

[2] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel et al., *Scikit-learn: Machine learning in Python*, *Journal of Machine Learning Research* **12** (2011) 2825.

[3] L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A. Mueller, O. Grisel et al., *API design for machine learning software: experiences from the scikit-learn project*, in *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pp. 108–122, 2013.

[4] V.D. Blondel, J.-L. Guillaume, R. Lambiotte and E. Lefebvre, *Fast unfolding of communities in large networks*, *Journal of Statistical Mechanics: Theory and Experiment* **2008** (2008) P10008.

[5] V.A. Traag, L. Waltman and N.J. van Eck, *From louvain to leiden: guaranteeing well-connected communities*, *Scientific Reports* **9** (2019) 1.

[6] J. Mauro and G. de Wasseige, *Searching for sub-populations within the gamma-ray solar flares catalog: a graph-based clustering analysis*, *PoS* **ICRC2023** (2023) 1292.

[7] R. Sibson, *SLINK: An optimally efficient algorithm for the single-link cluster method*, *The Computer Journal* **16** (1973) 30.

[8] R.A. Fisher, *The use of multiple measurements in taxonomic problems*, *Annals of Eugenics* **7** (1936) 179.

[9] S. Lloyd, *Least squares quantization in pcm*, *IEEE Transactions on Information Theory* **28** (1982) 129.

[10] R.J.G.B. Campello, D. Moulavi and J. Sander, *Density-based clustering based on hierarchical density estimates*, in *Advances in Knowledge Discovery and Data Mining*, (Berlin, Heidelberg), pp. 160–172, Springer Berlin Heidelberg, 2013, DOI.

[11] J.D. Banfield and A.E. Raftery, *Model-based gaussian and non-gaussian clustering*, *Biometrics* **49** (1993) 803.

[12] P.A. Evans et al., *Methods and results of an automatic analysis of a complete sample of Swift-XRT observations of GRBs*, *Mon. Not. Roy. Astron. Soc.* **397** (2009) 1177.

[13] A. Lien et al., *The Third Swift Burst Alert Telescope Gamma-Ray Burst Catalog*, *Astrophys. J.* **829** (2016) 7.

[14] Swift Science collaboration, *The Swift Gamma-Ray Burst Mission*, *Astrophys. J.* **611** (2004) 1005.