A Graph-Driven Approach to Complex Challenges: A Case Study on Multiobjective Stellar and Earth-Like Exoplanet Clustering

Matheus Silva[1], Adriano Pereira[2], Rafael do Carmo[1], Ariana Frühauf[3], and Michel Silva[1]

[1] [1] Federal Center for Technological Education of Minas Gerais (CEFET-MG),
Campus V,R. Álvares de Azevedo, 400, Bela Vista, 35503-822 ,
Divinópolis - MG, Brazil
matheus.silva@aluno.cefetmg.br
rafaelmarcelinocs@cefetmg.br
michel@cefetmg.br

[2] [2] Federal Institute of Maranhão (IFMA), Campus Imperatriz,
R. Dom Pedro II, S/N, Centro, 65900-090, Imperatriz - MA, Brazil
adriano.bezerra@ifma.edu.br

[3] [3] Adventist College of Minas Gerais (FADMINAS), Campus II,
R. Joaquim Gomes Guerra, 590, Kennedy, 37200-000 Lavras - MG, Brazil
arianafruhauf@gmail.com

Abstract. This study investigates the complex interplay between stellar metallicity, stellar mass, and the likelihood of hosting Earth-like planets. Recognizing that the formation of habitable planets is a multifactorial process, we adopt a novel graph-based approach that integrates Local-Sensitive Hashing (LSH) for efficient dimensionality reduction with modularity-based clustering to analyze a dataset of nearly 38,000 exoplanets. Our methodology enables us to identify distinct planetary communities and examine key host star properties. Comprehensive statistical analyses, including ANOVA, t-tests, and various correlation measures (Pearson, Spearman, Kendall, distance correlation, and mutual information), were employed to explore the relationships between stellar mass and planetary density. While our findings indicate that Sun-like stars—particularly those with slightly lower stellar mass—are more likely to host Earth-like planets, the observed correlations suggest that the influence of stellar mass is relatively weak, pointing to the involvement of additional factors in planetary formation. Moreover, the computational efficiency of our approach highlights its potential applicability in other fields that deal with high-dimensional, complex datasets, such as systems biology and social network analysis. Overall, our work contributes to a deeper understanding of the determinants of planetary habitability and underscores the need for more sophisticated multivariate models to capture the full complexity of planet formation.

Keywords: Earth-like Planets · Metallicity · Mass · Graph-Based Clustering · Computational Astrophysics

1 Introduction

The search for Earth-like exoplanets is one of the primary objectives of modern astrophysics, aiming to understand the conditions necessary for habitability and the prevalence of planets similar to Earth in the universe. With the increasing availability of exoplanetary data from missions such as Kepler, TESS, and various ground-based surveys, identifying and characterizing these planets has become a complex challenge that requires advanced computational techniques.

A key question in exoplanetary science is whether certain stellar properties influence the likelihood of hosting Earth-like planets. In particular, stellar mass and metallicity have been proposed as fundamental factors affecting planetary formation and composition. Higher metallicity is associated with an increased probability of forming rocky planets, while stellar mass determines the radiation and gravitational influences that shape planetary environments. Understanding these relationships is crucial for refining exoplanet detection models and improving our understanding of planetary system formation.

The formation of habitable planets can be viewed as a complex system, where multiple interacting variables govern the final planetary characteristics. In such systems, factors like stellar metallicity, mass, and other environmental conditions interact in non-linear ways, making the process inherently multivariate and dynamic. This complexity challenges traditional modeling approaches and necessitates the use of robust, computationally efficient algorithms.

To address these challenges, this study employs a novel approach that combines graph-based clustering with dimensionality reduction techniques. These methods are particularly effective in handling large, high-dimensional datasets—a common hurdle in exoplanet research. Graph-based clustering enables the visualization of intricate relationships between exoplanets by grouping planets that share similar characteristics, thereby facilitating the identification of patterns and potential Earth-like candidates.

One of the major computational challenges is efficiently comparing a multitude of exoplanetary properties (e.g., stellar mass, metallicity, orbital characteristics) across tens of thousands of entries. We employed Local-Sensitive Hashing (LSH) for dimensionality reduction, which accelerates processing by grouping similar planets into “buckets.” This approach significantly reduces the computational cost of pairwise comparisons, while preserving the similarity between objects—a critical advantage when dealing with massive datasets.

Once the similarity graph is constructed using these clustering methods, we analyze the stellar properties of the host stars, with a particular focus on metallicity and mass, to determine if Earth-like planets are preferentially hosted by Sun-like stars. Moreover, the methodology presented here not only refines exoplanet habitability models but also demonstrates potential applications in other fields that deal with complex, high-dimensional data.

2 Related Works

Several studies have explored the relationship between stellar properties and exoplanet formation using graph-based and statistical approaches, offering insights that align with our own. A recent study by [5] discusses how host star properties, including metallicity and stellar mass, correlate with the probability of forming Earth-like planets. This aligns with the finding that higher stellar metallicity tends to promote the formation of rocky planets, which are more likely to host conditions favorable for life. Similarly, [3] examines large-scale stellar surveys to identify exoplanet-hosting stars, reinforcing the importance of metallicity thresholds in rocky planet formation. In their work, [9] propose a new classification system for exoplanets based on host star attributes, employing machine learning and graph-based clustering methods similar to those used in our study. This approach supports the hypothesis that Sun-like stars are more likely to host planets with Earth-like conditions and introduces innovative clustering techniques that could improve our understanding of exoplanetary systems.

Additionally, [8] applies traditional clustering techniques, like K-means and hierarchical methods, to analyze exoplanetary data, offering insight into the relationship between planetary mass, orbital period, and eccentricity. Their study, which uses these methods to classify exoplanet populations, is instrumental in demonstrating the power of clustering for analyzing large astronomical datasets. Graph-based clustering methods have also emerged as powerful tools for identifying relationships within complex datasets. For instance, [4] investigates the stellar mass-metallicity relation, modeling the evolution of metallicity across different stellar populations and its effects on planetary composition, while [6] explores graph-based exoplanet classification and demonstrates that network analysis can reveal hidden structures in exoplanet datasets. By leveraging these techniques, researchers can gain a deeper understanding of the distribution of planetary systems and their potential for habitability.

Furthermore, [10] explores the use of deep learning in graph-based clustering methods, suggesting that such techniques can greatly enhance the accuracy of clustering in exoplanetary data analysis. Dimensionality reduction methods like Principal Component Analysis (PCA) and Local-Sensitive Hashing (LSH) have also been widely used in exoplanet research to reduce the complexity of large datasets. [5] discusses the role of dimensionality reduction in exoplanet classification, providing a framework for applying machine learning techniques to simplify the analysis process. In a related vein, [10] emphasizes the importance of data transformation techniques, such as PCA, in improving the performance of clustering algorithms applied to exoplanet datasets, reinforcing the choice of LSH as a dimensionality reduction technique. Finally, statistical methods have long played an important role: [8] highlights how t-tests and ANOVA can be applied to assess the statistical significance of various exoplanetary properties, and our own work extends these approaches by applying these tests to compare stellar mass and metallicity distributions between Earth-like and non-Earth-like planets. Machine learning pipelines have gained prominence in this area, as [5] shows how they can be used to classify exoplanets by integrating graph theory and clustering methods, while deep learning models have proven effective in enhancing the accuracy of exoplanet identification based on stellar properties.

3 Methodology

We began by obtaining our dataset from NASA's Exoplanet Archive [2], focusing on ten key features from the 288 available parameters to facilitate comparisons with Earth (Table 1). As a reference point for evaluating whether our approach could reliably cluster Earth-like systems, two data points were introduced: Earth itself and a slightly modified "artificial Earth-like planet" whose features closely resemble Earth but are not exact duplicates (Table 2). This design allows us to verify if our methods can consistently group planets with similar characteristics.

Table 1: Planetary and Stellar Characteristics Used for Comparison

Feature	**Description**
st_{mass}	Host star's mass
st_{met}	Host star's metallicity
st_{teff}	Host star's temperature
st_{radius}	Host star's radius
pl_{insol}	Planet's received radiation
pl_{eqt}	Planet's equilibrium temperature
pl_{orbper}	Planet's orbital period
$pl_{orbsmax}$	Planet's semi-major axis
$pl_{orbeccen}$	Planet's orbital eccentricity
pl_{dens}	Planet's density

Table 2: Comparison of Earth and Similar Earth Parameters

Feature	**Earth**	**Similar Earth**
Stellar Temperature (st_{teff})	5778 K	5778 K
Stellar Metallicity (st_{met})	0.0 [Fe/H]	0.01 [Fe/H]
Stellar Radius (st_{rad})	1.0 $R_\odot$	1.0 $R_\odot$
Stellar Mass (st_{mass})	1.0 $M_\odot$	1.0 $M_\odot$
Planet Insolation (pl_{insol})	1.0	1.02
Planet Equilibrium Temperature (pl_{eqt})	288 K	290 K
Orbital Period (pl_{orbper})	365.25 days	365.0 days
Semi-Major Axis ($pl_{orbsmax}$)	1.0 AU	1.01 AU
Orbital Eccentricity ($pl_{orbeccen}$)	0.0167	0.02
Planet Density (pl_{dens})	5.51 g/cm^3	5.5 g/cm^3

Data preprocessing involved a logarithmic transformation, $\log(x+1)$, to reduce the skewness caused by extreme values in certain parameters (e.g., very long orbital periods). We also performed a deduplication step, retaining only one record per unique planet name to avoid overrepresentation of any system. Following these steps, we benchmarked three approaches for dimensionality reduction and preliminary clustering: Local-Sensitive Hashing (LSH), K-means, and a KNN-based neighborhood graph. Over 10 executions, LSH achieved an average runtime of only 0.0067 seconds—dramatically faster than K-means (0.0410 seconds) and especially the KNN method (2.0243 seconds). Additionally, while the clustering similarity between LSH and K-means (Adjusted Rand Index, ARI, of 0.5067) indicates moderate agreement, the near-zero ARI values observed when comparing LSH and KNN (–0.0004) as well as K-means and KNN (–0.0003) suggest that the graph-based approach captures a markedly different structure. These findings underscore LSH as the preferred method due to its superior efficiency and its ability to reliably group planets with similar properties, serving as a robust foundation for further analyses.

With LSH selected, we first constructed a minimal graph by connecting only Earth to any planet in the same LSH bucket, thereby focusing our attention on the most Earth-like subset. We then used Kernel Density Estimation (KDE) as a non-parametric technique to visualize how the stellar properties of these candidate planets compared to the broader dataset. KDE produces a smooth estimate of the probability density function, giving us a clearer picture of whether Earth's bucket of planets shows distinct stellar mass or metallicity distributions compared to the remaining population. At this stage, no final conclusions were drawn—our goal was solely to observe if there were apparent shifts in distributions or potential clustering effects.

Next, we expanded the analysis to include a full similarity graph of all exoplanets in the dataset, weighting edges by $1/(\text{Euclidean distance})$. In order to identify potential communities, we employed a layout algorithm (Fruchterman–Reingold) to position the nodes and then applied modularity-based clustering. By partitioning the network into modules, we could assess whether Earth-like planets cluster together under more global conditions rather than just in the narrowly filtered LSH bucket.

Following the modularity partition, we performed statistical testing to check whether certain clusters had distinctive stellar characteristics. Specifically, we applied Analysis of Variance (ANOVA) to detect whether there were significant differences in stellar parameters (e.g., mass, metallicity) across multiple clusters. Where necessary, pairwise comparisons with t-tests were employed to examine whether an Earth-like cluster differed meaningfully from the others in average stellar mass or metallicity. ANOVA is well-suited for comparing the means of three or more groups, while t-tests allow more focused comparisons between two sets of data.

Finally, we investigated whether a linear relationship might exist between stellar mass (st_{mass}) and planetary density (pl_{dens}), particularly in Earth-like systems. To that end, we employed several correlation coefficients that capture linear or rank-based associations, including Pearson (for linear correlation), Spearman and Kendall (for monotonic relationships), and additional measures for potential non-linear dependencies (e.g., distance correlation). These metrics reveal whether increasing stellar mass correlates in a straightforward way with changes in planetary density, or whether the relationship is more complex.

In summary, the methodological pipeline comprised data cleansing, logarithmic scaling, LSH-based dimension reduction, graph building for candidate planets, KDE to gain an initial view of distribution differences, a global similarity graph for modularity clustering, and statistical testing (ANOVA, t-tests) to evaluate cluster distinctions. Lastly, we calculated correlation coefficients to determine how strongly stellar mass and planetary density might be related in those clusters. The numerical findings and more detailed interpretations of these analyses will be presented in the Results section.

4 Results

In this section, we present the findings of our graph-based clustering approach, following the methodological steps outlined earlier. By systematically comparing minimal Earth-like subsets and then expanding to a global exoplanet network, we demonstrate how high-dimensional data analysis and network-based clustering can help tackle the complexity of identifying Earth-like candidates.

4.1 LSH Filtering and Initial Visualization

After benchmarking different methods (LSH, K-means, and KNN), we selected LSH due to its computational efficiency. We first constructed a minimal graph connecting each planet in the same LSH bucket as Earth within it. Figure 1 shows this initial visualization, generated in Python using NetworkX and PyVis, where node size reflects stellar mass and node color indicates planetary density.

To gain an overview of how these Earth-like candidates differed from the broader dataset, we applied Kernel Density Estimation (KDE) to stellar mass and metallicity distributions. Figures 2 and 3 compare the distributions for planets in the Earth-like bucket versus the remaining population. KDE is a non-parametric method for estimating the probability density function of a continuous variable, offering a smoothed perspective relative to traditional histograms.

Overall, these KDE plots indicate that the LSH-filtered set of Earth-like planets tends to cluster around near-solar metallicity and stellar mass values, consistent with [3], which suggests that stars similar to the Sun might facilitate rocky planet formation.

4.2 Global Network Construction and Modularity Clustering

Although the initial LSH bucket highlights a core set of Earth-like candidates, we next constructed a global similarity graph of all exoplanets, weighting edges by 1/Euclidean Distance. The ForceAtlas2 and Fruchterman–Reingold layout was

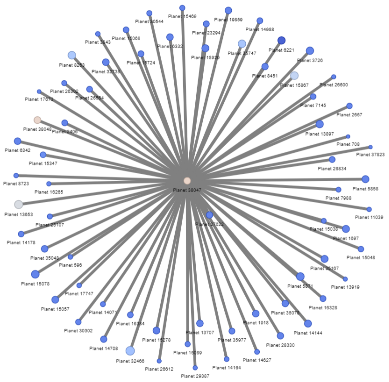


Fig. 1: Initial graph connecting Earth to similar planets identified via LSH. Node size represents stellar mass, and color represents planetary density.

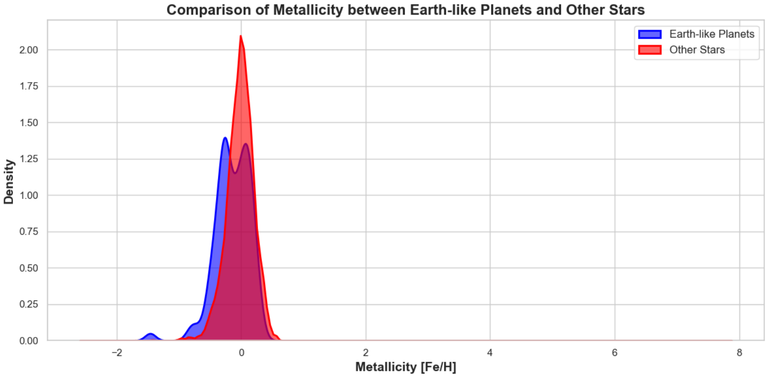


Fig. 2: KDE plot comparing stellar metallicity distribution between Earth-like planets (LSH bucket) and the broader stellar population.

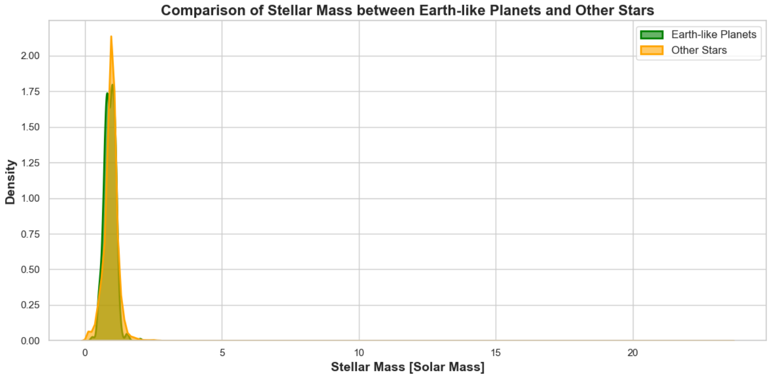


Fig. 3: KDE plot comparing stellar mass distribution between Earth-like planets (LSH bucket) and the broader stellar population.

then applied in Gephi to spatially organize nodes based on similarity, and a modularity algorithm was used to detect communities.

Figure 4 shows an intermediate network highlighting how Earth-like candidates link to Earth itself, while Figure 5 presents the final refined graph with a modularity-based color scheme.

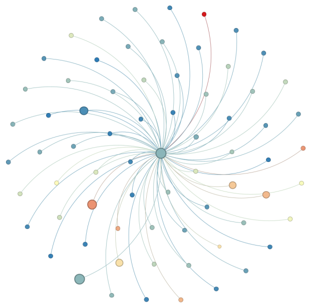


Fig. 4: Graph representation of Earth-like planet candidates after LSH filtering, showing direct connections to Earth. Nodes represent planets, edges represent feature-space similarity.

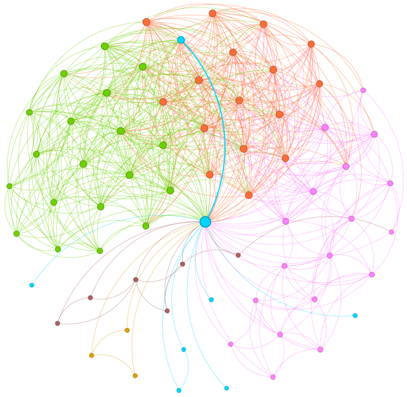
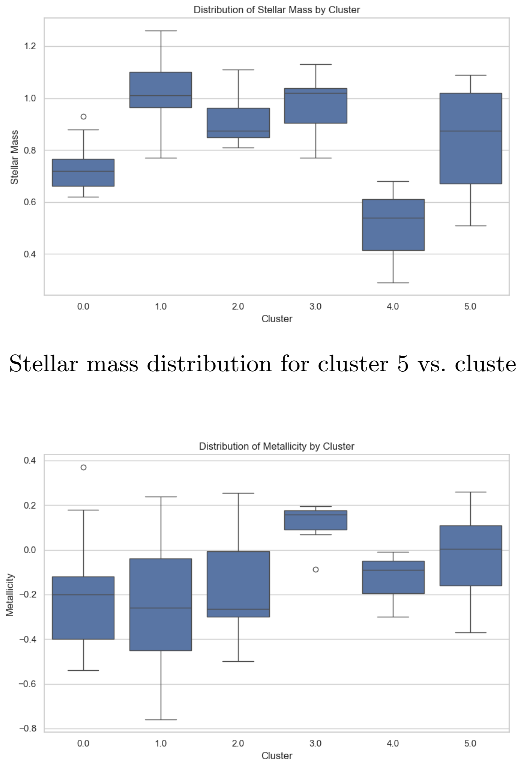


Fig. 5: Final exoplanet network, where edges are weighted by 1/Euclidean Distance. Colors indicate different modularity-based communities, revealing distinct planetary clusters.

A key outcome of this step is the identification of multiple communities, each representing exoplanets with mutually similar features. Notably, Earth and the artificial Earth-like planet occupy the same community, indicating that our similarity criteria successfully group genuine Earth analogs with slightly perturbed configurations.

4.3 Extended Statistical Analysis

To evaluate differences among the discovered clusters, we computed summary statistics and performed statistical tests. Figures 6(a) and (b) compare stellar mass and metallicity for Earth-like planets (cluster 5) versus the broader exoplanet population (clusters 0–4). The box plots illustrate that Earth-like planets typically orbit stars near solar mass ($\approx 1\,M_\odot$) and metallicity ([Fe/H] ≈ 0).



(a) Stellar mass distribution for cluster 5 vs. clusters 0–4.

(b) Stellar metallicity distribution for cluster 5 vs. clusters 0–4.

Fig. 6: (a) and (b) show how Earth-like planets in cluster 5 compare with other clusters in terms of mass and metallicity.

Two-sample t-tests confirm that cluster 5 differs significantly ($p < 0.05$) in key stellar parameters compared to most other clusters. This finding echoes the proposition that Sun-like stars may present more conducive environments for Earth-like planet formation [5].

Finally, we tested whether a linear or non-linear relationship exists between stellar mass (st_{mass}) and planetary density (pl_{dens}). Table 3 summarizes multiple correlation coefficients (Pearson, Spearman, Kendall), polynomial regression fits, and measures such as Distance Correlation and Mutual Information.

Table 3: Correlation and Regression Results for st_{mass} vs. pl_{dens} (Earth-like Systems)

Method	**Result**
Pearson Correlation	-0.1665 (p = 0.509)
Spearman Correlation	-0.1718 (p = 0.495)
Kendall Correlation	-0.1063 (p = 0.542)
Polynomial Regression (Degree 2)	$R^2 = 0.0406$
Polynomial Regression (Degree 3)	$R^2 = 0.2171$
LOESS Smoothing	Correlation = -0.4122
Distance Correlation	0.4119
Mutual Information	0.0010

The overall correlation values are low (< 0.2 for Pearson, Spearman, Kendall), indicating weak linear or monotonic relationships. Although Distance Correlation suggests the possibility of a minor non-linear association, polynomial models fail to produce a meaningful fit (R^2 below 0.22). Thus, we infer that stellar mass alone does not strongly predict planetary density and that additional factors—such as planetary composition or system architecture—may have greater influence.

4.4 Summary of Results

In summary, our results highlight that:

- Sun-like stars ($\approx 1\,M_\odot$, [Fe/H] ≈ 0) are more likely to host planets identified as “Earth-like,” affirming prior findings [3].
- Graph-based modularity clustering effectively isolates Earth analog communities, including our artificially perturbed Earth twin.
- Host star properties alone (especially stellar mass) exhibit weak direct correlations with planetary density, suggesting that additional variables shape the habitability and composition of exoplanets.

These findings collectively underscore the value of high-dimensional analysis and network-based clustering in understanding complex exoplanetary data. Our approach supports the theme of *"Making Complex Systems Tractable through Computational Science"* by providing robust tools for discovering and characterizing potentially habitable worlds.

5 Discussion

Analyzing nearly **38,000 planets** required efficient methods for high-dimensional data reduction and clustering. We evaluated:

- **Local-Sensitive Hashing (LSH):** Chosen for its efficiency in handling large datasets.
- **K-means Clustering:** Performed well but was computationally more expensive.
- **K-Nearest Neighbors (KNN):** Computationally prohibitive (≈2.27s per query, compared to 0.017s for LSH).

Takeaways:

- LSH **significantly outperformed K-means and KNN** in execution time, making it the optimal choice.
- Future research may integrate **semi-supervised machine learning** to enhance clustering performance.

5.1 Broader Applications and Future Work

Our graph-based approach has potential applications beyond astrophysics, including:

- **Biology:** Detecting genetic relationships in high-dimensional data.
- **Social Networks:** Identifying hidden communities in online interactions.
- **Finance:** Modeling complex market correlations using network-based clustering.

Future Directions:

- Expanding the model to consider **stellar age, activity, and disk properties**.
- Comparing graph-based clustering with **deep learning techniques** for exoplanet classification.
- Applying this methodology to upcoming missions, such as **TESS and JWST**.

6 Conclusion

In conclusion, our study introduced a novel graph-based approach to investigate the relationship between host star properties and the presence of Earth-like exoplanets. By constructing a similarity graph and applying modularity-based clustering, we successfully identified distinct planetary communities within a dataset of nearly 38,000 exoplanets. Our comprehensive statistical analysis, which included ANOVA and two-sample t-tests, revealed significant differences in stellar properties among the clusters. In particular, while both stellar mass and metallicity vary across clusters, stellar mass emerged as a more discriminative factor in distinguishing the Earth-like cluster (cluster 5) from the others.

These results support the hypothesis that Sun-like stars—especially those with slightly lower stellar mass—may offer more favorable conditions for the formation or detection of Earth-like planets. Moreover, the integration of Local-Sensitive Hashing (LSH) for dimensionality reduction with graph-based clustering not only enhanced computational efficiency but also provided a robust framework that can be adapted to other complex systems.

Overall, our work contributes to the broader effort of refining exoplanet detection and habitability models by offering new insights into the interplay between stellar and planetary properties. Future research should aim to incorporate additional parameters and explore more sophisticated multivariate models to further elucidate the complex processes governing planetary system formation.

References

1. Matheus Silva, “Stellar Insights: A repository for exoplanetary data analysis and graph clustering,” GitHub Repository, 2025. Available: `https://github.com/Matheus-Emanue123/StellarInsights`.
2. NASA Exoplanet Archive, “Planetery Systems Table,” 2025. Available: `https://exoplanetarchive.ipac.caltech.edu/cgi-bin/TblView/nph-tblView?app=ExoTbls&config=PS`.
3. S. Stock et al., “Giant planet occurrence as a function of stellar mass and metallicity: A unified view from CARMENES and other radial velocity surveys,” *Astronomy & Astrophysics*, vol. 674, A75, 2023. Available: `https://www.aanda.org/articles/aa/full_html/2023/06/aa45387-22/aa45387-22.html`.
4. M. J. Dickens et al., “The stellar mass-metallicity relation of exoplanet host stars: Insights from APOGEE,” 2018. Available: `https://arxiv.org/abs/1810.08928`.
5. R. Heller et al., “The CARMENES search for exoplanets around M dwarfs: A network approach to planetary system architectures,” *Astronomy & Astrophysics*, vol. 674, A25, 2023. Available: `https://arxiv.org/html/2411.17358v1`.
6. H. A. G. Colmenares, “Graph-based exoplanet classification,” *University of São Paulo (USP)*, 2018. Available: `https://colmenares2018.exoplanetarydata`.
7. Gephi Consortium, “Gephi: The Open Graph Viz Platform,” 2025. Available: `https://gephi.org/`.
8. I.-G. Jiang, L.-C. Yeh, W.-L. Hung, and M.-S. Yang, “Data analysis on the extra-solar planets using robust clustering,” *arXiv:astro-ph/0610695v1*, 2006. Available: `https://arxiv.org/abs/astro-ph/0610695v1`.
9. D. J. Armstrong et al., “Graph-based community detection of exoplanetary systems,” *Monthly Notices of the Royal Astronomical Society*, vol. 532, no. 2, pp. 2832–2845, 2024. Available: `https://academic.oup.com/mnras/article/532/2/2832/7700710`.
10. A. P. C. et al., “Deep learning in graph clustering for exoplanet classification,” *arXiv:2407.09055v1*, 2024. Available: `https://arxiv.org/abs/2407.09055v1`.