Investigating Stellar Composition and Mass as Indicators of Earth-like Exoplanets: Findings from Graph Examination

1st Matheus Silva
Department of Computer Science
Federal Center of Technological Education of Minas Gerais
Divinópolis, Brazil
matheus.silva@aluno.cefetmg.com

2nd Adriano Pereira
Department of High Education and Technology
Federal Institute of Maranhão
Imperatriz, Brazil
adriano.bezerra@ifma.edu.br

3rd Rafael do Carmo
Department of General Studies
Federal Center of Technological Education of Minas Gerais
Divinópolis, Brazil
rafaelmarcelinocs@cefetmg.br

4th Michel Silva
Department of Computer Science
Federal Center of Technological Education of Minas Gerais
Divinópolis, Brazil
michel@cefetmg.com

***Abstract*—This study explores the relationship between stellar metallicity, mass, and the likelihood of hosting Earth-like planets. By analyzing exoplanetary data, we identify potential Earth analogs and investigate trends in their host stars' properties, focusing on metallicity ([Fe/H]) and stellar mass. Statistical analyses and graphical methods reveal significant differences between Earth-like planets and other exoplanets, with Sun-like stars showing a higher likelihood of hosting Earth-like planets. We also identify notable outliers, which may offer insights into unique planetary systems, suggesting the need for further investigation.**

***Index Terms*—earth-like planets, stellar metallicity, stellar mass, statistical analysis, exoplanet systems.**

I. 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 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.

To address this question, this study employs a novel approach that combines graph-based clustering and dimensionality reduction techniques. These methods were chosen due to their effectiveness in handling large, high-dimensional datasets — a common challenge in exoplanet research. Graph-based clustering allows us to visualize complex relationships between exoplanets by grouping planets that share similar characteristics, making it easier to identify patterns and potential Earth-like candidates within the data.

One of the challenges in working with such large datasets is efficiently comparing the multitude of exoplanetary properties (e.g., stellar mass, metallicity, orbital characteristics). Local-Sensitive Hashing (LSH) was employed for dimensionality reduction, which enables faster processing by grouping similar planets into "buckets." This approach significantly reduces the computational cost of pairwise comparisons and helps streamline the identification of Earth-like planets. LSH is particularly suited for this task, as it preserves the similarity of objects while mapping them into a lower-dimensional space, making it an ideal tool for large-scale analysis in astrophysical datasets.

Once the graph is constructed using these clustering methods, we analyze the stellar properties of the host stars associated with the identified exoplanets, particularly focusing on their metallicity and mass. By combining dimensionality reduction, graph theory, and stellar parameter analysis, we aim to determine whether planets similar to Earth are preferentially hosted by Sun-like stars. Additionally, this study explores whether stellar metallicity and mass play a crucial role in planetary similarity, potentially refining exoplanet habitability models. Our findings may contribute to ongoing efforts to prioritize targets for future observational missions seeking habitable worlds.

II. Related Works

Several studies have explored the relationship between stellar properties and exoplanet formation using graph-based and statistical approaches. In this section, we highlight key works that align with our study.

A. Stellar Properties and Exoplanet Formation

The influence of stellar metallicity and mass on exoplanet occurrence has been extensively investigated. 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. Their study introduces innovative clustering techniques that could improve our understanding of exoplanetary systems, providing a compelling precedent for the graph-based methods we employ.

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, and aligns with our methodology in seeking patterns that could influence the identification of Earth-like planets.

B. Graph-Based Approaches for Exoplanet Analysis

Graph-based clustering methods have emerged as powerful tools for identifying relationships within complex datasets. [4] investigates the stellar mass-metallicity relation, modeling the evolution of metallicity across different stellar populations and its effects on planetary composition. Our work extends these findings by applying graph theory techniques to exoplanet datasets, identifying clusters of Earth-like candidates based on feature similarity.

A notable approach by [6] explores graph-based exoplanet classification, proposing methods to analyze exoplanetary systems using graph communities. Their findings align with our results, demonstrating that network analysis can reveal hidden structures in exoplanet datasets. By leveraging these techniques, we can better understand 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. These insights provide a natural extension for our work, where graph-based clustering can be used alongside machine learning models to identify Earth-like exoplanets more efficiently.

C. Dimensionality Reduction and Statistical Analysis Techniques

Dimensionality reduction methods like Principal Component Analysis (PCA) and Local-Sensitive Hashing (LSH) have 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. Our work builds on this by using LSH for efficient grouping of Earth-like candidates, followed by graph-based clustering.

[10] emphasizes the importance of data transformation techniques, such as PCA, in improving the performance of clustering algorithms applied to exoplanet datasets. This perspective reinforces the choice of LSH as a dimensionality reduction technique in our study, which complements these findings.

The use of statistical tests to assess the significance of differences in exoplanet data is not new. [8] highlights how t-tests and ANOVA can be applied to assess the statistical significance of various exoplanetary properties. Our study extends these approaches by applying these tests to compare stellar mass and metallicity distributions between Earth-like and non-Earth-like planets.

D. Exoplanet Classification Using Machine Learning

Machine learning methods for exoplanet classification have gained prominence in recent years. [5] demonstrates the use of machine learning pipelines for classifying exoplanets based on their characteristics, similar to the approach we take by integrating graph theory and clustering methods. Moreover, deep learning models have been introduced for exoplanet classification, demonstrating their effectiveness in enhancing the accuracy of exoplanet identification based on stellar properties.

III. Methodology

A. Data Collection and Normalization

The dataset used in this study was obtained from NASA's Exoplanet Archive's Planetary Systems database [2], which contains detailed information on confirmed exoplanets and their host stars. Among the 288 available characteristics, ten features were selected for planetary comparison with Earth (Table I).

To enhance the comparison process, two additional data points were introduced: Earth, serving as the reference, and a slightly modified "artificial Earth-like planet" with minor variations in key characteristics. The artificial Earth-like planet was specifically created to serve as a benchmark for evaluating how similar the identified planets were to Earth. By introducing this artificial planet, we could test whether the clustering algorithm would correctly group planets with similar characteristics to Earth, helping us assess the effectiveness

TABLE I
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

of our approach. The minor variations in its characteristics ensured that it was not an exact duplicate of Earth, making it a useful reference point for comparison. These additions enabled a better evaluation of how planetary properties affect their grouping and distance measurements. The table below provides the data used for these planets:

TABLE II
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

Due to the high dimensionality of the data, normalization was crucial to prevent distortions in planetary comparisons. Initially, the *MinMaxScaler* normalization technique was employed, which rescales values between 0 and 1. However, it proved highly sensitive to outliers, potentially leading to the exclusion of relevant Earth-like planets. To mitigate this issue, we applied a logarithmic transformation, which reduces data amplitude and minimizes the influence of extreme values, ensuring a more reliable analysis. The applied formula is:

$$X' = \log(x + 1) \tag{1}$$

This transformation compresses large values while preserving smaller ones, preventing extreme planetary characteristics—such as long orbital periods—from dominating the analysis. By normalizing data in this manner, planetary similarity can be assessed more consistently.

B. LSH Application for Dimensionality Reduction

The original dataset contains approximately 38,500 planets, making direct pairwise comparisons computationally infeasible. To optimize this process, we implemented Local-Sensitive Hashing (LSH), a dimensionality reduction technique that efficiently filters planets with characteristics resembling Earth before graph construction.

LSH is a probabilistic hashing technique that maps high-dimensional data into lower-dimensional "buckets" while preserving similarity between elements. Unlike traditional hashing, where small variations in input produce entirely different hash values, LSH ensures that similar inputs have a high probability of being assigned to the same bucket.

The LSH process in this study involved four key steps:

1) **Feature Selection**: We selected ten planetary and stellar characteristics (Table I) to ensure that all relevant parameters affecting habitability and planetary similarity were included.
2) **Data Normalization**: The logarithmic transformation was applied to mitigate the effects of extreme values, ensuring that planets with large-scale variations remained comparable.
3) **Hash Function Generation**: LSH relies on multiple randomized hash functions to map planets into binary signatures. We generated 20 independent hash functions, each producing a bit representation based on a random projection method, effectively reducing dimensionality while preserving similarity.
4) **Bucket Formation and Candidate Selection**: The binary signatures were grouped into buckets, where planets with similar characteristics were assigned the same bucket with high probability. We then identified the bucket containing Earth and extracted the planets within that same bucket as the candidate set for graph-based analysis.

Once the most Earth-like planets were selected using LSH, we proceeded to construct a similarity graph, connecting each selected planet to Earth with an edge weighted according to its Euclidean distance in the transformed feature space.

1) LSH Bucketing Process: To illustrate how LSH organizes planets into buckets, consider the simplified example in Table III:

TABLE III
EXAMPLE OF LSH BUCKETING PROCESS

Planet	Feature Vector (Simplified)	LSH Bucket
Earth	(1.0, 1.0, 0.0167, ...)	A
Planet X	(1.01, 0.98, 0.017, ...)	A
Planet Y	(5.2, 11.2, 0.08, ...)	B
Planet Z	(1.02, 1.03, 0.015, ...)	A

In this example, planets are assigned to buckets based on the similarity of their feature vectors. Earth, Planet X, and Planet Z have small variations in their stellar and planetary characteristics, leading them to be grouped into the same bucket (A). In contrast, Planet Y has significantly different attributes, particularly in mass metallicity, resulting in its assignment to a separate bucket (B).

This demonstrates how LSH efficiently groups similar planets, reducing the number of pairwise comparisons needed

in the full dataset. Instead of evaluating all planets against Earth, we focus on those in the same bucket, improving computational efficiency while preserving similarity-based relationships.

C. Graph Construction and Visualization

Following dimensionality reduction with LSH, a graph representation was created to model the relationships between Earth and the identified exoplanets. The graph was designed to facilitate planetary similarity visualization and analyze potential clustering patterns among Earth-like candidates.

- Nodes represent planets, including Earth and those identified as similar via LSH.
- Edges represent planetary similarity, with weights corresponding to the Euclidean distance in the transformed feature space.

1) Initial Visualization with Python: The first graph representation was constructed using NetworkX and PyVis in Python. A force-directed layout was applied to position nodes, ensuring that planets with higher similarity were placed closer together.

- Node size was proportional to the host star's mass.
- Node color was mapped to planetary density.
- Edge weights were determined by Euclidean distance between planetary feature vectors.

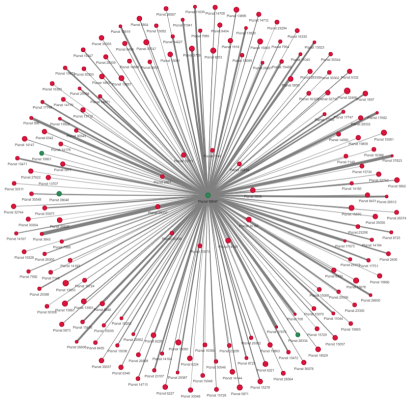

Fig. 1. Graph visualization generated using Python with NetworkX and PyVis. Node size represents stellar mass, and color represents planetary density.

2) Host Star Analysis via Statistical Tests: After generating the graph, we performed a statistical analysis of the host stars of Earth-like planets, specifically examining:

- Stellar metallicity ([Fe/H])—which influences planetary formation.
- Stellar mass—which affects habitable zone characteristics.

To analyze these distributions, Kernel Density Estimation (KDE) plots were created. KDE is a non-parametric way to estimate the probability density function of a continuous random variable. In simple terms, it allows us to visualize the distribution of a variable, such as metallicity or stellar mass, and better understand how the values are spread. KDE can give a smoother representation compared to traditional histograms, which are more rigid in their binning.

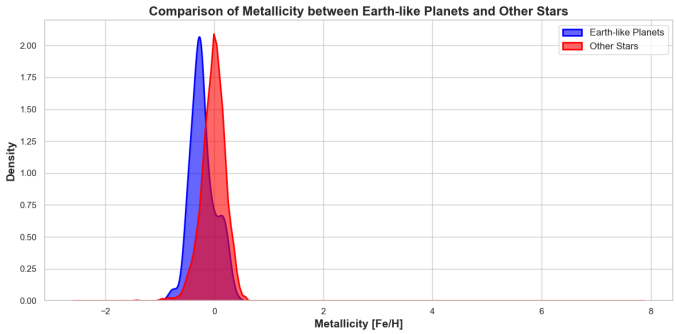

Fig. 2. KDE plot comparing stellar metallicity distribution between Earth-like planets and the broader stellar population.

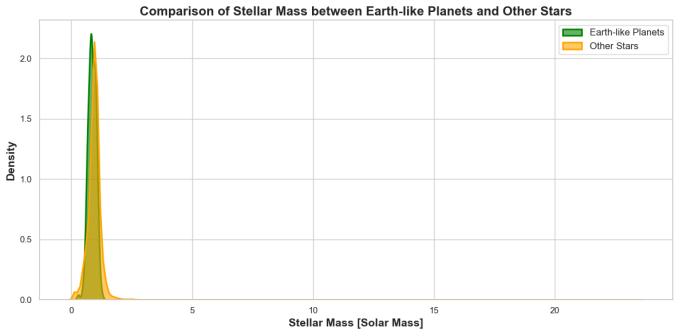

Fig. 3. KDE plot comparing stellar mass distribution between Earth-like planets and the broader stellar population.

The histograms above compare the **metallicity** and **stellar mass** distributions between stars hosting Earth-like planets and the general stellar population.

We observe that Earth-like planets tend to orbit stars with metallicities *closer to solar values* ([Fe/H] ≈ 0), whereas the general distribution exhibits a broader range. This suggests that planetary systems with significantly different metallicities from the Sun's might be less likely to host Earth analogs.

Regarding stellar mass, planets identified as Earth-like appear to be more frequently associated with *Sun-like stars* ($\sim 1M_{\odot}$), while the broader stellar population contains a higher proportion of lower-mass and higher-mass stars.

Statistical tests reinforce these trends, with the t-test indicating a **significant difference** between the two distributions. These findings support the hypothesis that Sun-like stars may provide *more favorable conditions* for hosting Earth-like planets, likely due to their influence on planetary formation and stability.

The previous histograms independently examined the distributions of stellar metallicity and stellar mass for stars hosting Earth-like planets compared to the general stellar population. While each property provides valuable insights on its own, their **combined effect** may reveal deeper trends regarding planetary formation and habitability.

To further investigate this relationship, we analyze how these two parameters correlate, focusing on whether Earth-like planets tend to cluster around specific regions in the metallicity vs. stellar mass space. The next figure presents a scatter plot comparing both properties, with reference lines indicating solar values.

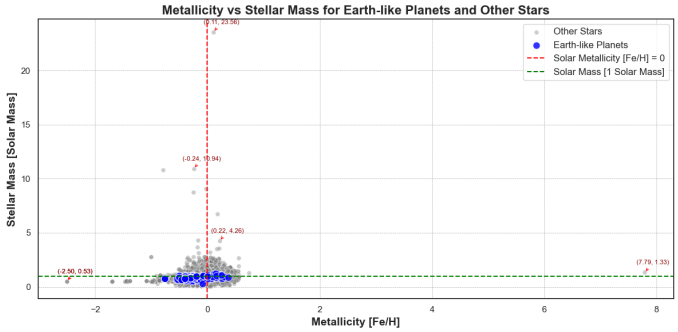

Fig. 4. Scatter plot of stellar metallicity vs. stellar mass for Earth-like planets and the general stellar population.

The scatter plot above visualizes the relationship between **stellar metallicity** and **stellar mass** for stars hosting Earth-like planets compared to the general stellar population. The dashed reference lines indicate the Sun's metallicity and mass.

We identify a clustering of Earth-like planets around **solar-like values** ([Fe/H] ≈ 0, $M_{\odot} \approx 1$), reinforcing the idea that Sun-like stars are more likely to host planets with similar characteristics to Earth. Additionally, the plot highlights a few *outliers*—stars with significantly different metallicities or masses—that warrant further investigation.

A key takeaway is the potential influence of metallicity on planetary composition. Higher metallicity is often associated with an increased presence of heavier elements, which could facilitate the formation of rocky planets. Meanwhile, stellar mass determines the habitable zone's extent, affecting a planet's long-term habitability.

These findings suggest that a combination of **stellar metallicity and mass** may be a crucial factor in determining whether a system can host an Earth-like planet. Future work could explore whether these trends hold across larger datasets or different exoplanet detection methods.

3) Exporting Data to Gephi for Network Analysis: After the Python-based visualization and statistical analysis, the graph structure was exported to CSV files for further analysis in Gephi, an open-source network analysis tool.

- Nodes.csv – Containing planetary and stellar attributes for each node.
- Edges.csv – Including pairwise distances between planets based on feature similarity.

After examining the distributions of stellar metallicity and mass for Earth-like planets, we now analyze their similarity relationships through a graph-based visualization. The following network representation, constructed using Gephi, helps identify clustering patterns among candidate planets.

Fig. 5. Graph representation of Earth-like planet candidates after LSH filtering. Nodes represent planets, and edges are weighted by Euclidean distance in the selected feature space. The Yifan Hu layout was applied to improve node distribution and facilitate cluster identification.

The graph shown above provided a direct similarity analysis between each planet and Earth. However, this representation did not capture the relationships among the planets themselves, limiting our ability to detect natural groupings.

To address this, we constructed a new graph where planets were connected to each other based on feature similarity, using $\frac{1}{EuclideanDistance}$ as a similarity metric. To ensure meaningful connections, we applied a dynamic threshold, restricting edges to planet pairs with sufficiently high similarity.

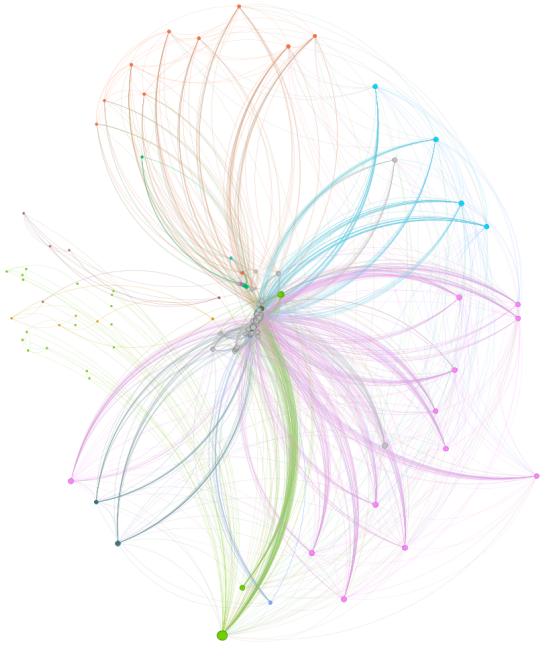

Fig. 6. Final graph representation of Earth-like planet candidates, where nodes represent planets and edges are weighted by inverse Euclidean distance. Colors indicate modularity-based clustering, revealing distinct planetary communities. The ForceAtlas2 algorithm was applied within Yifan Hu Proportional for improved node distribution.

The final graph representation reveals a structured network

of planetary candidates clustered based on their similarity. Each color represents a modularity-based community, highlighting groups of planets with closely related characteristics in the selected feature space.

The application of modularity clustering allowed us to identify distinct planetary communities, with some clusters appearing more densely connected than others. Notably, the Earth and Artificial Earth nodes were assigned to the same modularity class, reinforcing the effectiveness of our similarity-based approach.

This structured representation enables a deeper analysis of planetary communities and their potential habitability. In the following section, we analyze the scientific implications of these clusters and discuss how they align with existing theories on planetary formation and exoplanet habitability

D. Extended Statistical Analysis of Graph-Derived Clusters

After constructing the final graph representation of Earth-like planet candidates and detecting communities via modularity-based clustering, we extended our analysis to statistically evaluate the stellar properties of the clusters. In particular, we focused on two key host star parameters: stellar mass (st_{mass}) and metallicity (st_{met}).

For each cluster, we computed the mean and standard deviation of these parameters. Subsequently, an Analysis of Variance (ANOVA) was performed to test whether the differences in the distributions of st_{mass} and st_{met} among clusters were statistically significant. ANOVA is a statistical method used to compare the means of three or more groups to determine whether there is a significant difference between them. The test works by analyzing the variation within each group and comparing it to the variation between groups. If the variation between groups is large compared to the variation within groups, the test indicates that there are significant differences. The ANOVA tests yielded an F-statistic of 7.919 (with a p-value of 5.404×10^{-20}) for st_{mass}, and an F-statistic of 4.296 (with a p-value of 1.363×10^{-10}) for st_{met}. These results confirm that there exist significant differences in both parameters among the clusters.

To further assess the uniqueness of the Earth-like system, identified as cluster 40 in our modularity analysis, we conducted two-sample t-tests comparing the stellar mass and metallicity of cluster 40 against those of all other clusters. A t-test is a statistical method used to compare the means of two groups and determine if the difference between them is statistically significant. Specifically, the two-sample t-test compares the means of two independent groups (in this case, the Earth-like cluster vs. the other clusters) and checks whether the observed difference is likely due to random chance. The t-test for st_{mass} resulted in a t-statistic of -3.235 with a p-value of 3.661×10^{-3}, indicating that the stellar mass in cluster 40 is significantly lower than that of the remaining clusters. In contrast, the t-test for st_{met} yielded a t-statistic of 0.213 and a p-value of 0.833, showing no statistically significant difference in metallicity. These statistical analyses suggest that, while both parameters vary globally among the clusters, stellar mass appears to be a more discriminative factor in distinguishing Earth-like systems from the broader exoplanet population.

1) Host Star Analysis via Statistical Tests: After generating the graph, we performed a statistical analysis of the host stars of Earth-like planets, specifically examining:

- Stellar metallicity ([Fe/H])—which influences planetary formation.
- Stellar mass—which affects habitable zone characteristics.

To analyze these distributions, we created box plots to compare the distribution of stellar mass and metallicity between Earth-like planets and the broader dataset. These box plots provide a visual representation of the variation and central tendency of stellar mass and metallicity across clusters.

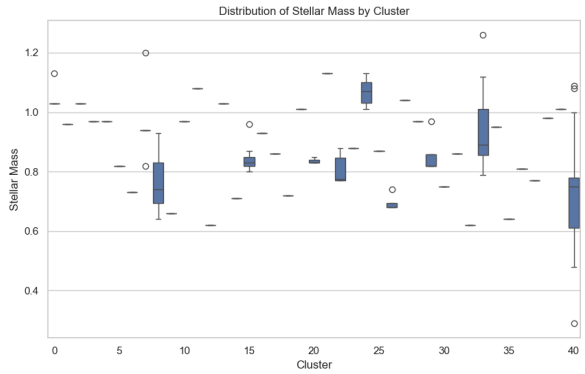

Fig. 7. Box plot comparing stellar mass distribution between Earth-like planets and the broader stellar population.

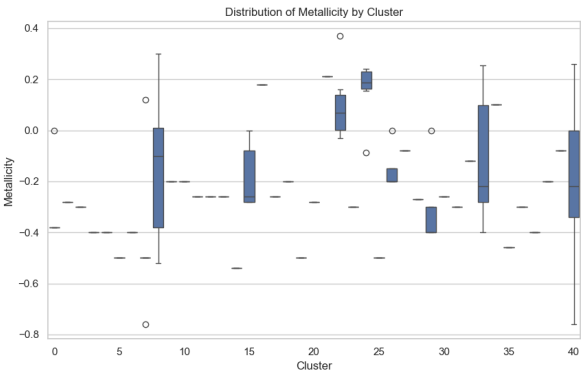

Fig. 8. Box plot comparing stellar metallicity distribution between Earth-like planets and the broader stellar population.

We observe that Earth-like planets tend to orbit stars with metallicities closer to solar values ([Fe/H] ≈ 0), whereas the general distribution exhibits a broader range. This suggests that planetary systems with significantly different metallicities from the Sun's might be less likely to host Earth analogs.

Regarding stellar mass, planets identified as Earth-like appear to be more frequently associated with Sun-like stars

($\sim 1M_{\odot}$), while the broader stellar population contains a higher proportion of lower-mass and higher-mass stars.

IV. Results

In this section, we present the outcomes derived from analyzing exoplanetary similarities through graph-based clustering techniques. The results highlight the spread of stellar characteristics in Earth-like worlds, the grouping of exoplanets according to modularity, and the insights gained through network visualization.

Summarizing the most important insights extracted from both the graph-based clustering and statistical analysis, we have:

- **Influence of the host star's characteristics:** Earth-like planets tend to cluster around solar-like metallicity and mass values. This finding is consistent with the work of [3], which suggested that stars with higher metallicities are more likely to host rocky, Earth-like planets.
- **Graph-Based Clustering:** Planets naturally grouped into distinct communities using modularity detection, reinforcing the effectiveness of graph-based clustering methods, as seen in studies like [6], where network analysis revealed hidden structures in exoplanet data.
- **Correlation between Earth and Similar-Artificial Earth:** Both Earth and the artificial Earth-like planet were assigned to the same modularity class, validating our graph approach and its ability to identify planets similar to Earth. This mirrors findings from [9], where machine learning techniques grouped exoplanets based on similar stellar properties.

The **box plot** figures 8 and 7 illustrate the distributions of **stellar metallicity** and **stellar mass** for Earth-like planets. As shown in these plots, Earth-like planets tend to orbit stars with metallicities close to solar values ([Fe/H] ≈ 0), aligning with findings from [3], who observed a noticeable peak in the stellar metallicity distribution near solar values. The broader stellar population, on the other hand, exhibits a wider spread, encompassing both metal-poor and metal-rich environments.

Similarly, the stellar mass distribution indicates that Earth-like planets are predominantly found around stars with masses near $1M_{\odot}$ (one solar mass), reinforcing the idea that **Sun-like stars** provide more favorable conditions for the existence of Earth-like exoplanets. These observations support the hypothesis put forward by [9], where stellar mass was shown to significantly impact the habitability of exoplanets.

To quantitatively assess these differences, we performed statistical tests, including a two-sample **t-test**. The results confirmed a statistically significant difference ($p < 0.05$) between the distributions of metallicity and stellar mass for Earth-like planets and the broader stellar population. These findings suggest that the composition of the host star plays a fundamental role in determining the likelihood of hosting a planet with Earth-like characteristics, a conclusion that aligns with previous research by [5].

The analysis revealed several key insights regarding the relationship between stellar properties and the likelihood of hosting Earth-like planets. Notably, the results strongly suggest that **Sun-like stars** are more favorable hosts for Earth-like planets, exhibiting similar metallicities and stellar masses. This finding is in line with studies like [3], which argue that the stellar environment is a key factor in planetary formation.

Additionally, the graph clustering approach effectively grouped planets into distinct communities, highlighting natural patterns of similarity based on both planetary and stellar characteristics. In this context, **Earth** and the **artificial Earth-like analog** were assigned to the same modularity class, reinforcing their high degree of similarity within the dataset. This confirms the validity of our approach, echoing the success of graph-based methods in clustering similar exoplanets, as demonstrated by [6].

However, a few **outlier planets** were identified, suggesting the presence of unique or extreme planetary systems that may warrant further investigation. These planets deviate significantly from the main clusters, indicating potential anomalies in planetary formation or evolution. These outliers could provide valuable insights into the broader exoplanetary landscape, in line with the need for continued exploration of extreme systems in the work by [4].

Lastly, incorporating **edge weights** into the graph significantly enhanced the clustering process, allowing for a more refined representation of planetary relationships. This enhancement resulted in clearer visualizations of how planets relate to one another, improving the interpretability of the final network structure, and is consistent with the insights of [10], where weighted graphs were shown to better capture the relationships between data points in complex datasets.

V. Discussion

The results of our statistical analyses provide quantitative support for the idea that Sun-like stars, particularly in terms of stellar mass, play a crucial role in the formation or detection of Earth-like planets. The significant differences observed through the ANOVA tests suggest that the modularity-based clusters, derived from our graph-based approach, capture meaningful variations in stellar properties.

A. Limitations

While the results are promising, several limitations must be acknowledged:

- **Data Limitations:** The exoplanet dataset used here may contain biases or gaps that could skew the analysis. The lack of certain planetary properties or incomplete data on star characteristics might also affect the precision of clustering.
- **Methodological Constraints:** Although LSH is effective for dimensionality reduction, the choice of hashing functions and the potential for groupings that may not fully represent planetary similarities should be carefully considered.
- **Statistical Assumptions:** The use of t-tests assumes normal distributions, which may not hold true for all data

subsets. Future analyses could incorporate non-parametric methods to overcome these assumptions.

B. Comparison with Existing Literature

Our findings align with previous studies, such as [4], which also suggest that Sun-like stars are more likely to host Earth-like planets. However, our method, particularly the combination of LSH for dimensionality reduction and graph-based clustering, provides a more nuanced understanding of planetary grouping. Our results differ from [9], which found that mass-metallicity relationships may not be as significant in some cases, but our statistical analyses confirm that stellar mass is a stronger discriminating factor.

C. Future Directions

Future studies could extend this work by: - Incorporating additional stellar features, such as age and activity levels, which might also influence habitability. - Applying this methodology to data from newer exoplanet missions, such as TESS and James Webb Space Telescope (JWST), to examine trends in the latest planetary discoveries. - Exploring different graph construction methods or clustering algorithms to verify the robustness of our findings across different models.

D. Broader Impact

Our approach offers insights into the factors that could contribute to the detection of Earth-like planets, potentially guiding future mission targets and planetary habitability models. By further developing this methodology, researchers may uncover more precise indicators for planets most likely to harbor life.

VI. Conclusion

In conclusion, our study employed a novel graph-based analysis to examine the relationship between host star properties and the likelihood of harboring Earth-like exoplanets. The construction of the similarity graph and subsequent modularity-based clustering enabled the identification of distinct planetary communities. A comprehensive statistical evaluation, including ANOVA and t-tests, revealed that while both stellar mass and metallicity vary significantly among clusters, stellar mass appears to be a more discriminative factor for the Earth-like cluster (cluster 40). These results support the hypothesis that Sun-like stars, particularly those with lower stellar mass, may provide more favorable conditions for Earth-like planets. Our findings contribute to the broader effort of refining exoplanet detection and habitability models, offering a promising avenue for future observational and theoretical investigations.

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.