Universitat Rovira i Virgili
Escola Tècnica Superior d'Enginyeria

Community Structure of Complex Networks

Master in Health Data Science — MHEDAS

Subject: Complex Networks

Author: Sofía González Estrada, Ravneet-Rahul Sandhu Singh

Date: November 23, 2025

UNIVERSITAT DE VIC
UNIVERSITAT CENTRAL
DE CATALUNYA

Universitat de Girona

UGA
Université
Grenoble Alpes

UNIVERSITAT DE BARCELONA

Universitat de Lleida

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH

UNIVERSITAT ROVIRA i VIRGILI

BIOINFORMATICS BARCELONA

UAB
Universitat Autònoma de Barcelona

eit Health

Co-funded by the European Union

Table of Contents

1. Introduction

The selected network to study is the Leeds Butterfly Dataset [1], which consists of images of various butterfly species.

1.1 Leeds Butterfly Dataset

The Leeds Butterfly Dataset contains images of 10 different butterfly species, with a total of 832 images. Each species has a varying number of images, ranging from 50 to 100 images per species. In **Table 1.1**, we provide a detailed description of each species included in the dataset.

Table 1.1: Species included in the Leeds Butterfly Dataset with descriptions.

Scientific Name	Common Name	Description
Danaus plexippus	Monarch	89-102 mm. Large with long forewings. Bright, burnt-orange with black veins and margins sprinkled with white dots.
Heliconius charitonius	Zebra Longwing	76-78 mm. Wings long and narrow. Jet-black above, banded with lemon-yellow. Bases of wings have crimson spots beneath.
Heliconius erato	Crimson-patched Longwing	76-86 mm. Wings long and rounded. Black above, crossed on forewing by a broad crimson patch, and on hindwing by a narrow yellow line.
Junonia coenia	Common Buckeye	51-63 mm. Tawny-brown to dark brown. Features distinctive large eyespots on each wing (black, yellow-rimmed with iridescent blue/lilac).
Lycaena phlaeas	American Copper	22-28 mm. Forewing bright copper with dark spots; hindwing dark brown with copper margin. Undersides mostly grayish with black dots.
Nymphalis antiopa	Mourning Cloak	73-86 mm. Large with ragged margins. Rich brownish-maroon above with a cream-yellow band bordered by brilliant blue spots.
Papilio cresphontes	Giant Swallowtail	86-140 mm. Very large with spoon-shaped tails. Dark brownish-black with broad bands of yellow spots.
Pieris rapae	Cabbage White	32-48 mm. Milk-white above with charcoal tips and black submarginal spots. Below, pale to bright mustard-yellow.
Vanessa atalanta	Red Admiral	44-57 mm. Black with orange-red to vermilion bars across forewing and hindwing border. White spots at the wing tip.
Vanessa cardui	Painted Lady	51-57 mm. Salmon-orange with black blotches and black-patterned margins. Forewings have black tips with clear white spots.

In addition to the images, the dataset also provides the segmentation masks for each butterfly, which can be useful for task like image segmentation and object detection. In **Figure 1.1**, we show some sample images from the dataset and their corresponding segmentation masks.

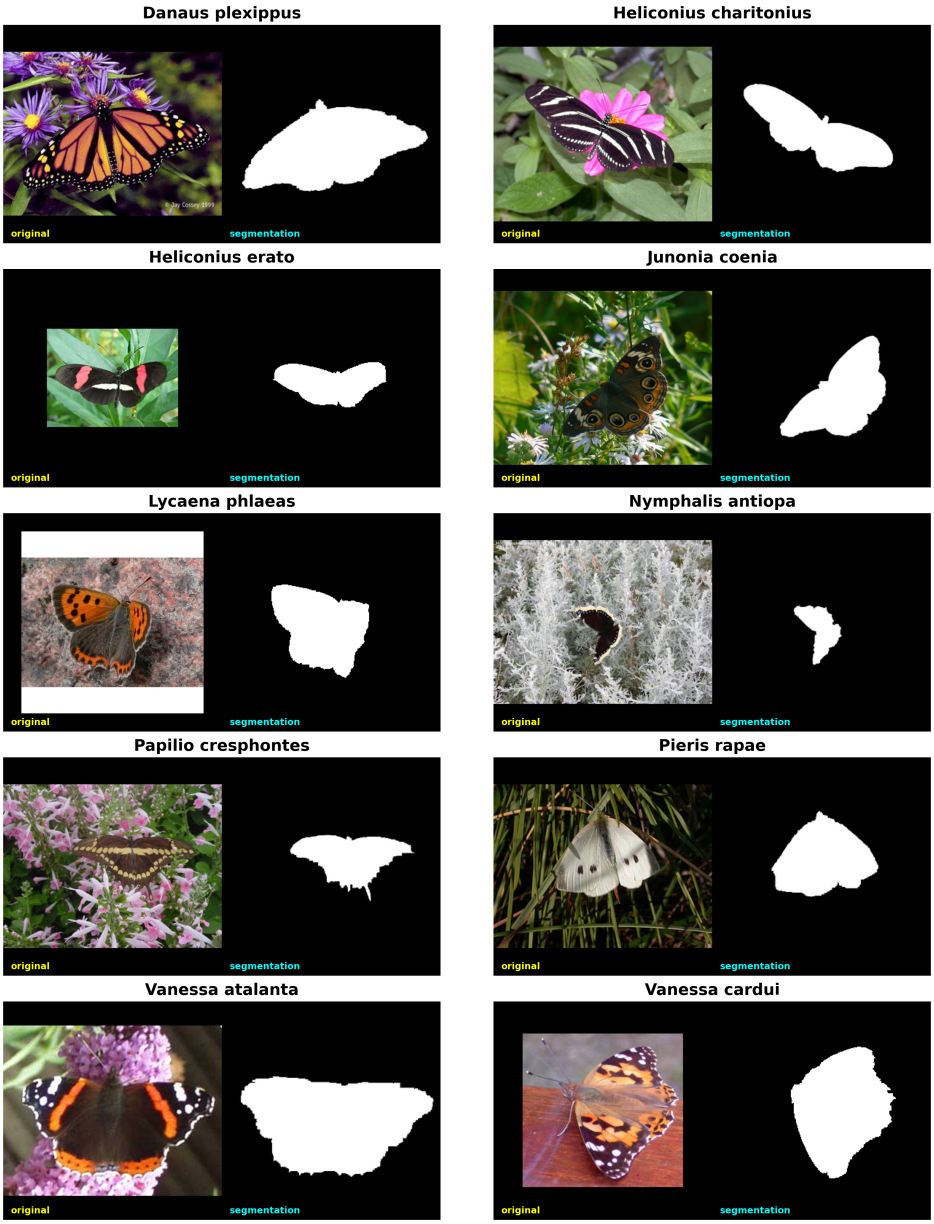

Figure 1.1: Sample images from the Leeds Butterfly Dataset.

1.2 Similarity Network

In the context of this project, the dataset also provides a similarity network based on visual likeness between the butterfly species. The network is represented as a graph, where nodes represent individual butterfly images, and edges represent visual similarities between the images. The edges are weighted based on the degree of similarity, with higher weights indicating greater similarity. Therefore, it is safe to assume that there will be communities formed by images of the same or similar species, as they are more likely to be visually close to each other. So, ultimately, our goal is to find these communities and see how well they correspond to the actual species labels.

2. Structural Analysis

Even though is not required for this project, we have performed a macroscopic and microscopic analysis of the Leeds Butterfly Dataset similarity network to gain a deeper understanding of its structure and properties.

2.1 Macroscopic

The Leeds Butterfly Dataset similarity network consists of 832 nodes and 86,528 edges. The average degree of the network is 208, with a minimum degree of 33 and a maximum degree of 530. This indicates that some nodes are highly connected, while others have relatively few connections. The network has an average clustering coefficient of 0.5954, suggesting a moderate level of clustering among nodes. In the context of this network, this means that if two butterfly images are both visually similar to a third image, there is a high probability that they are also visually similar to each other. The assortativity of the network is 0.2239, indicating a slight tendency for nodes to connect with others that have similar degrees. The average path length of the network is 1.8044, and the diameter is 4, indicating that the network is relatively compact, with most nodes being reachable from any other node within a few steps. In **Table 2.1**, we summarize the key macroscopic metrics of the Leeds Butterfly Dataset similarity network.

Table 2.1: Summary of Leeds Butterfly Dataset network metrics.

Metric	Value
Nodes	832
Edges	86,528
Average Degree	208.0000
Min. Degree	33
Max. Degree	530
Avg. Clustering Coefficient	0.5954
Assortativity	0.2239
Avg. Path Length	1.8044
Diameter	4

If we further explore the degree distribution of the network in **Figure 2.1**, we observe that rather than following a scale-free power-law distribution, the network exhibits a dense, homogeneous structure. Unlike scale-free networks where most nodes have a low degree, here the degrees are normally distributed around a high average of 208, with a minimum degree of 33, indicating that every image shares strong visual similarities with many others. Furthermore, the CCDF plot displays a curve rather than the straight line characteristic of scale-free power laws, confirming that the topology is driven by broad visual overlap between species rather than by a few central hubs.

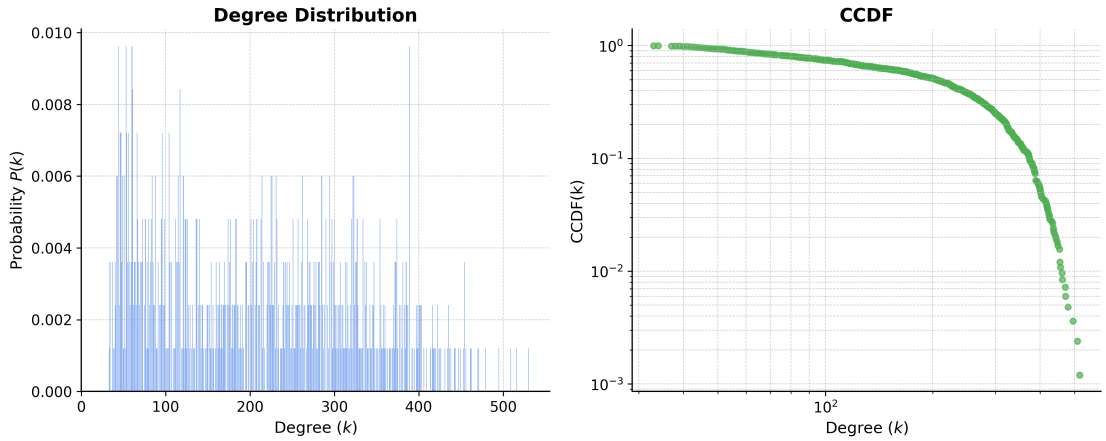

Figure 2.1: Degree distribution of the Leeds Butterfly Dataset similarity network.

2.2 Microscopic

To analyze the microscopic structure of the Leeds Butterfly Dataset similarity network, we computed three centrality measures: degree centrality, betweenness centrality, and eigenvector centrality. As shown in **Table 2.2**, the nodes with the highest degree centrality are largely the same as those with the highest eigenvector centrality. Notably, these central nodes are dominated by Species 10 (*Vanessa cardui*), Species 9 (*Vanessa atalanta*), and Species 5 (*Lycaena phlaeas*). This indicates that these species form the dense core of the network, likely due to visual features that make them visually similar to a large portion of the dataset. However, the nodes with the highest betweenness centrality differ entirely from the core group, belonging to Species 3 (*Heliconius erato*), Species 8 (*Pieris rapae*), and Species 1 (*Danaus plexippus*). This suggests that while *Vanessa* and *Lycaena* images act as archetypes with high connectivity, specific images from *Heliconius* and *Pieris* play a unique topological role as bridges. In the context of this dataset, these bridge nodes likely represent images with changing visual features that serve as links connecting different species clusters.

Table 2.2: Top 5 central nodes according to centrality measures.

Rank	Degree Centrality		Betweenness Centrality		Eigenvector Centrality	
	Node (Species)	Score	Node (Species)	Score	Node (Species)	Score
1	829 (10)	0.6378	234 (3)	0.0173	742 (9)	0.0726
2	742 (9)	0.6209	638 (8)	0.0112	391 (5)	0.0712
3	391 (5)	0.6125	64 (1)	0.0103	829 (10)	0.0708
4	374 (5)	0.5957	178 (3)	0.0099	374 (5)	0.0669
5	179 (3)	0.5764	571 (7)	0.0096	345 (5)	0.0667

As part of our microscopic analysis, we have also plotted the network in **Figure 2.2**. This kamada-kawai layout positions nodes based on their connections, aiming to distribute nodes evenly.

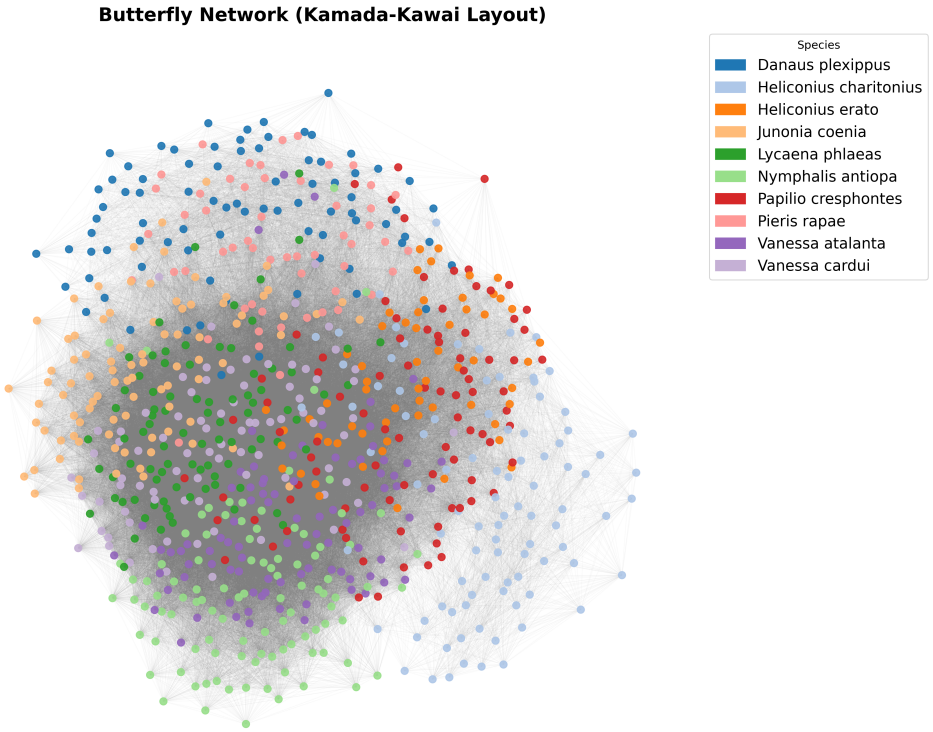

Figure 2.2: Kamada-Kawai layout of the Leeds Butterfly Dataset similarity network.

3. Community Detection

In this section, we explore various community detection algorithms applied to the Leeds Butterfly Dataset similarity network. In particular, we focus on three widely used methods: the Louvain algorithm, the Stochastic Block Model (SBM), and Infomap.

3.1 Modularity-Based Method: Louvain Algorithm

For the modularity-based approach, we employed the weighted Louvain algorithm. We ruled out the Greedy algorithm because, as discussed in class, it is significantly slower on dense networks. While the Leiden algorithm offers refinements over Louvain, it is not native to NetworkX. Therefore, we selected Louvain for its balance of efficiency and ease of implementation. As shown in **Figure 3.1**, Louvain identified 7 distinct communities. This under-segmentation shows how algorithm grouped centrally located into shared clusters. While the ones remaining in the surroundings were segregated into their own communities.

3.2 Stochastic Block Model (SBM)

The Stochastic Block Model (SBM) was employed taking into account the weighted edges of the network. Different runs of the algorithm were performed, and the one scoring minimum description length was retained. This resulted in a partition of 43 blocks (**Figure 3.2**), indicating a significant over-segmentation compared to the biological ground truth.

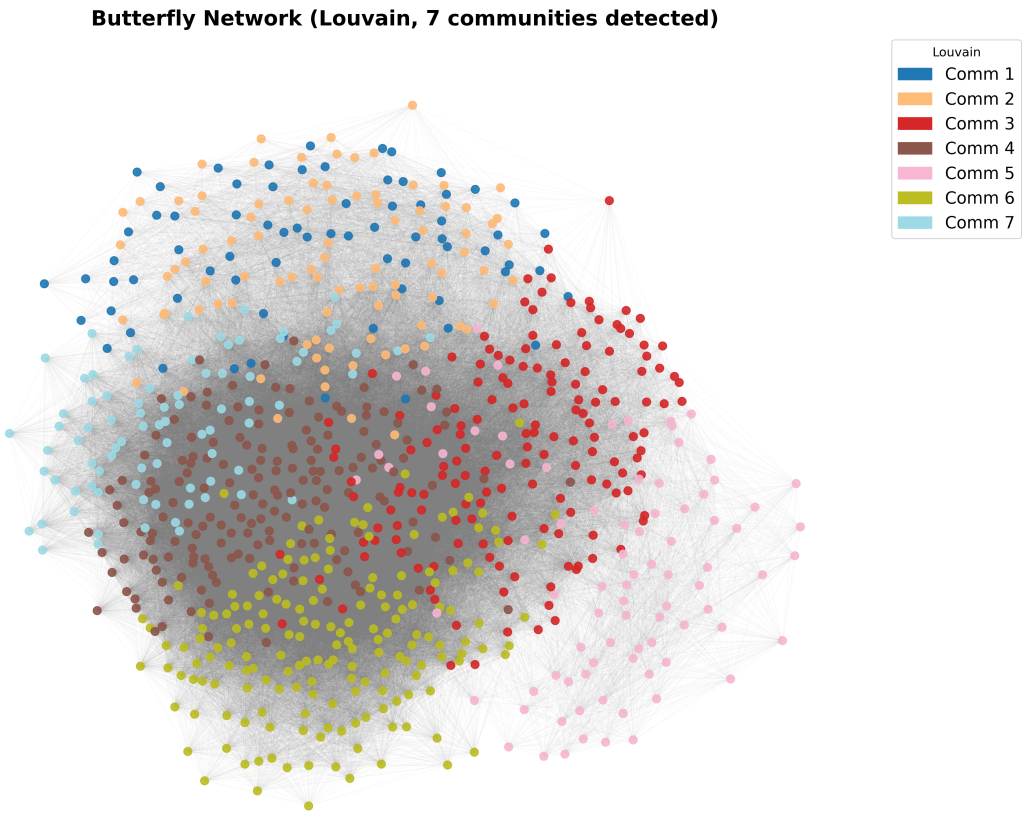

Figure 3.1: Community structure detected by the Louvain algorithm.

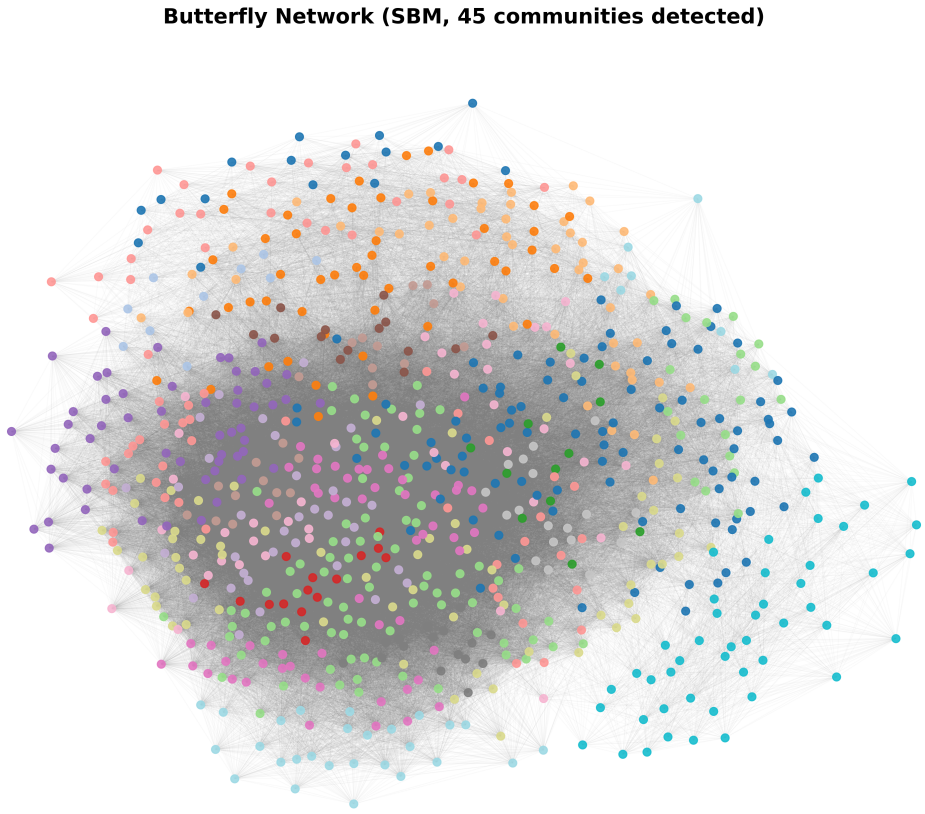

Figure 3.2: Community structure detected by the Stochastic Block Model (SBM).

The reason why we think this is happening is that, unlike algorithms that simply maximize density, SBM looks for structural equivalence. Given the extremely high density of our network, the algorithm likely detected subtle statistical patterns driven by visual noise rather than just species identity. In its attempt to minimize the description length, SBM interpreted these small visual variations as distinct blocks, effectively overfitting the model to the noise rather than capturing the biological traits.

3.3 Infomap

Finally, we applied the Infomap algorithm, which was configured to take into account the weights of the edges. **Figure 3.3** shows that the resulting partition consists of a total of 10 communities, which aligns directly with the number of butterfly species. Consequently, Infomap appears to be the most effective algorithm for capturing the underlying biological structure of the dataset. Although a closer inspection reveals that some nodes are misclassified into neighboring communities, the overall performance is quite satisfactory if we are only looking to recover the species labels.

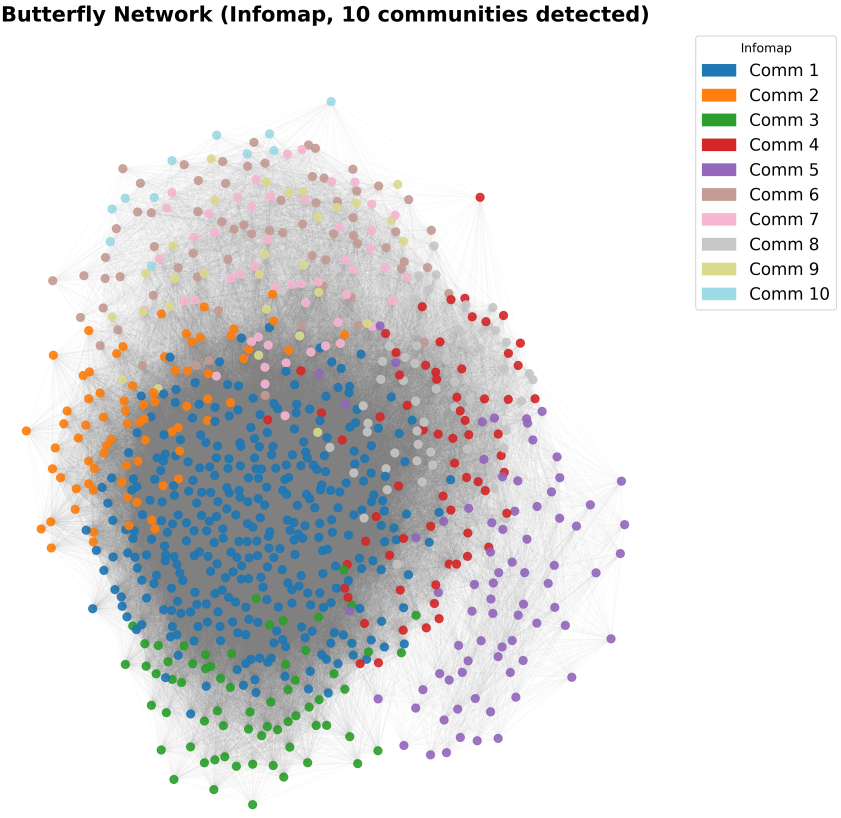

Figure 3.3: Community structure detected by the Infomap algorithm.

4. Discussion

Based on our results and looking back at our goals, we can say that SBM is definitely not the most suitable algorithm for this specific dataset, as it largely over-segments the network. On the other hand, while Infomap gives us the closest approximation to the

biological ground truth, it is not really what we are looking for in the sense that we gain no additional insight about the data other than just the species labels. Therefore, if we were to choose an algorithm that balances between capturing the biological structure and providing meaningful insights, the Louvain method would be our pick.

4.1 Louvain Confusion Matrix

In order to understand the communities generated by the Louvain algorithm, we generated the following confusion matrix. In **Figure 4.1**, we compare the 7 communities identified by Louvain (x-axis) against the 10 actual biological species (y-axis). The images on the x-axis correspond to the representative image of each community, defined as the node with the highest degree centrality within that specific cluster.

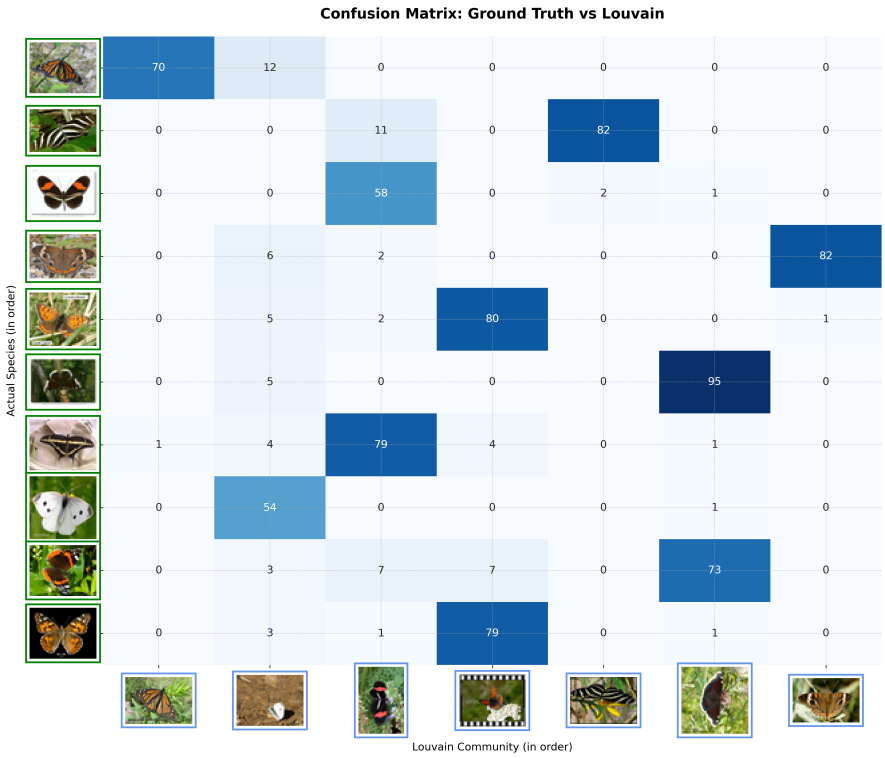

Figure 4.1: Confusion matrix comparing Louvain communities to ground truth species labels.

So, without any of us being experts in butterflies, we can still make some guesses on how the algorithm grouped the species based on visual similarities. First of all, we must clarify that even though the website mentioned that the edges are based on visual similarities, we do not know what this specifically means. Did they consider the color, the shape, the patterns, or a combination of all these features? So, before starting our deliberation, we decided to base our analysis mainly on colors and shape.

Community 1 is dominated by Species 1 (*Danaus plexippus*), and the colors and shape are quite unique, so it makes sense that it is segregated. Community 2 is mainly Species 8 (*Pieris rapae*), which is also quite unique in shape and the only white one, so again, it makes sense.

Now, Community 3 is where things start to get interesting, as it is a mix of Species 3 and 7 (*Heliconius erato* and *Papilio cresphontes*). Looking at the representative image (Y-axis), it is clearly Species 3, so why do we think Species 7 got grouped here? Well, we feel it's because of the horizontal white stripe that goes from wing to wing, which is common to both species.

Community 4 is mainly with Species 5 and 10 (*Lycaena phlaeas* and *Vanessa cardui*). Even though they're different species, this is a logically understandable behavior from the algorithm because, to the naked eye, both look very similar in shape and color, as they share close shades of orange and brown.

Moving on, Community 6 is mainly a mix of Species 6 and 9 (*Nymphalis antiopa* and *Vanessa atalanta*). At first, these two might seem very different, because Species 6 is characterized by cream-yellow bands and Species 9 has very distinctive orange bands. However, individuals of the two species have dark colored wings with much brighter bands on both the forewing and the hindwing, which explains why the model merged them into one community. Finally, Community 7 is mainly made up of Species 4 (*Junonia coenia*), a fairly easy species to distinguish thanks to the unique large eyespots on their wings.

4.2 Species Distribution

The species distribution across the communities is presented in **Figure 4.2**, which displays the data as a stacked bar chart. The percentage label centered on the largest segment of each bar indicates the purity of that community relative to its dominant species. For example, Community 1 is composed at 98.6% of Species 1 (*Danaus plexippus*). It also shows how segmented certain communities are. For instance, even though Community 2 is mainly composed of Species 8, it contains small traces of Species 1, 4, 5, 6, 7, 9, and 10, making it the most diverse community in terms of species variety.

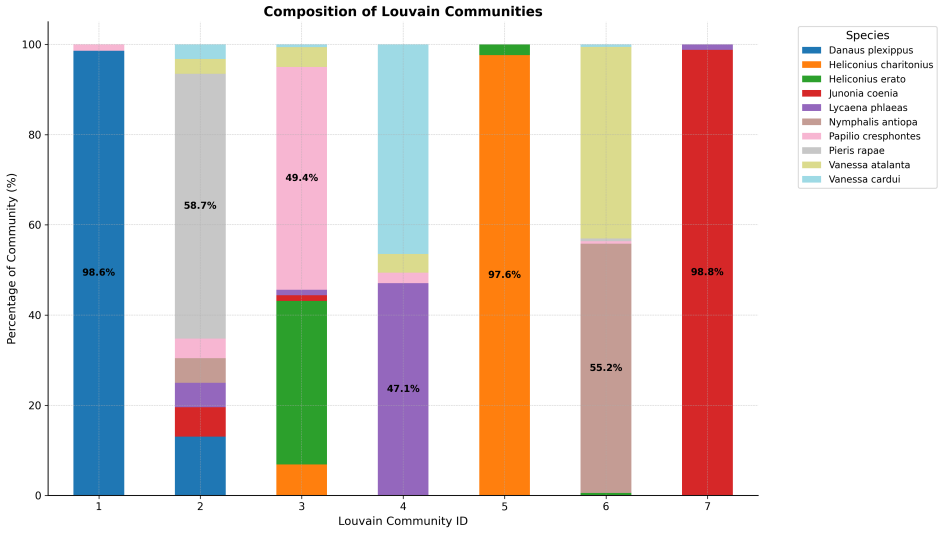

Figure 4.2: Stacked bar chart showing species distribution across Louvain communities.

We can also analyze the species distribution from another perspective. In **Table 4.1**, we provide a detailed breakdown of the percentage of each butterfly species assigned to the different Louvain communities. For instance, we can see that 95.1% of Species 3 and 88.8% of Species 7 were grouped into Community 3. This means that even though the Louvain algorithm merged the two species, it still isolated the individuals in them from the other 10 species.

Table 4.1: Percentage of each butterfly species assigned to the different Louvain communities.

Species	Community Distribution
Danaus plexippus	85.4% → Community 1 14.6% → Community 2
Heliconius charitonius	88.2% → Community 5 11.8% → Community 3
Heliconius erato	95.1% → Community 3 3.3% → Community 5 1.6% → Community 6
Junonia coenia	91.1% → Community 7 6.7% → Community 2 2.2% → Community 3
Lycaena phlaeas	90.9% → Community 4 5.7% → Community 2 2.3% → Community 3 1.1% → Community 7
Nymphalis antiopa	95.0% → Community 6 5.0% → Community 2
Papilio cresphontes	88.8% → Community 3 4.5% → Community 2 4.5% → Community 4 1.1% → Community 1 1.1% → Community 6
Pieris rapae	98.2% → Community 2 1.8% → Community 6
Vanessa atalanta	81.1% → Community 6 7.8% → Community 4 7.8% → Community 3 3.3% → Community 2
Vanessa cardui	94.0% → Community 4 3.6% → Community 2 1.2% → Community 3 1.2% → Community 6

4.3 Dendogram Analysis

Finally, we explored the hierarchical relationships within the communities identified by the Louvain algorithm. Instead of analyzing every cluster, we focused on the com-

munity with the highest diversity of species, which is Community 2. As noted earlier, this group is primarily dominated by Species 8 (*Pieris rapae*) but also contains traces of Species 1, 4, 5, 6, 7, 9, and 10. **Figure 4.3** presents the dendrogram for this specific community, allowing us to visualize the internal structure.

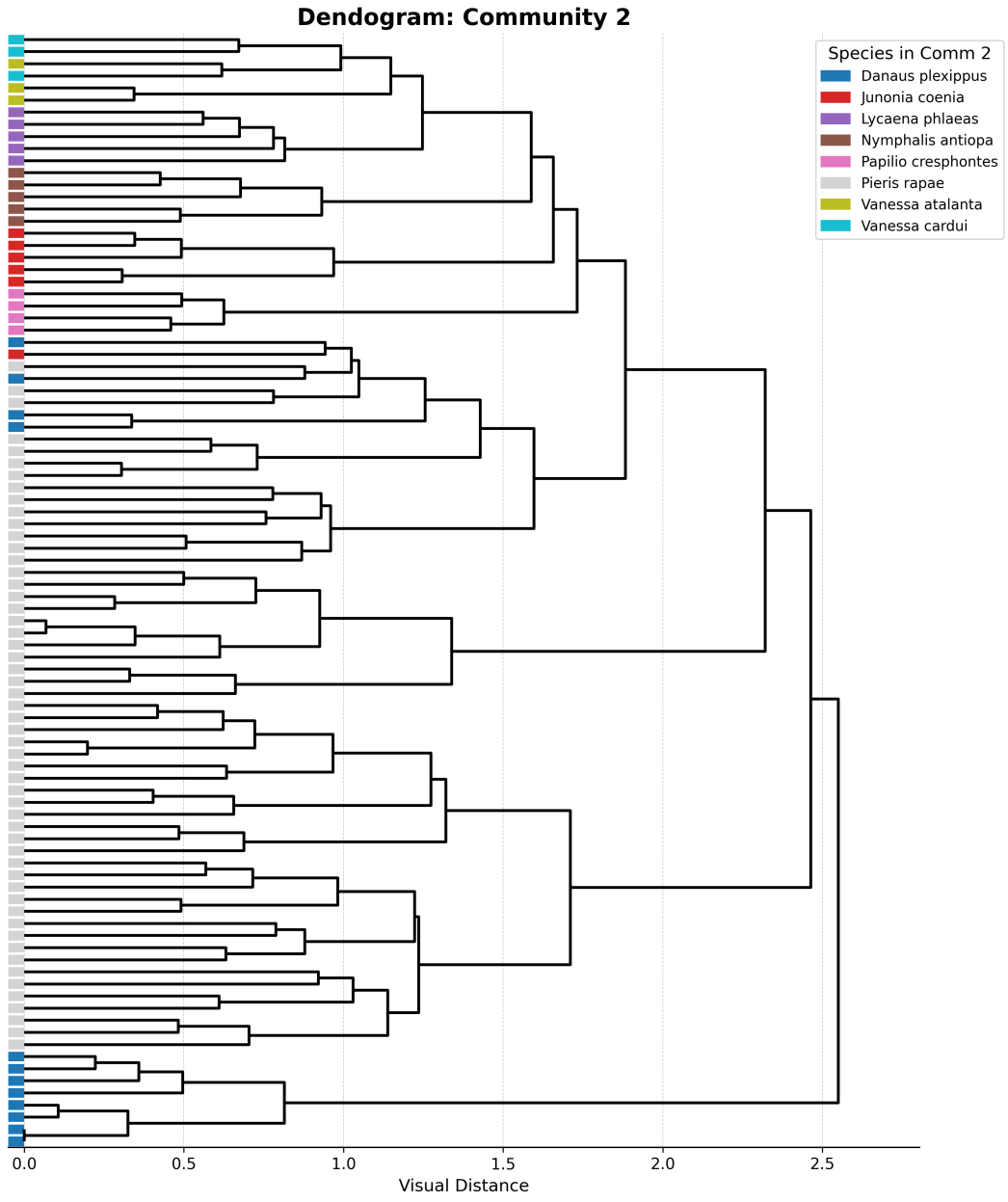

Figure 4.3: Dendrogram of Louvain Community 4 showing hierarchical clustering of nodes.

The visualization reveals a structured separation: the dominant Species 8 forms a solid, cohesive block at the bottom of the tree (gray bars), while the other species are not merely random noise but form their own branches at the top. Notably, Species 1 (*Danaus plexippus*) forms a sub-block distinct from the scattered remnants of other species. This indicates that the algorithm did not blindly confuse these species with the white butterflies. Instead, it identified them as a distinct visual sub-group that was structurally closer to the white butterflies than to their original biological groups.

5. Conclusions

In this project, we analyzed the Leeds Butterfly Dataset similarity network, where our goal was satisfactorily accomplished by the Louvain method, which provided a reasonable partition of the network into 7 communities. The results suggest that the Leeds Butterfly Dataset can be reduced into fewer categories based on visual phenotype, a conclusion supported by three analytical perspectives.

First, the confusion matrix revealed that the algorithm's deviations from the biological ground truth were not random errors but logical regroupings based on shared textures and shape patterns. Second, the species distribution analysis demonstrated that community purity is clearly linked to unique visual patterns; species defined by unique shapes (such as the circular eyespots of *Junonia coenia* or the stripes of *Heliconius charitonius*) formed the most isolated and pure clusters, while those defined merely by color were more prone to merging. Finally, the hierarchical dendrogram provided the structural context for the most heterogeneous group (Community 2), proving that even the *noise* was organized. The algorithm successfully identified distinct sub-groups of outliers clustering them together before merging them with the white butterflies based on general visual similarity. Therefore, these results confirm that community detection algorithms can successfully uncover a robust visual taxonomy that operates independently of, yet parallel to, biological classification.

References

[1] Josiah Wang, Katja Markert, and Mark Everingham. Learning models for object recognition from natural language descriptions. In *Proceedings of the British Machine Vision Conference*, 2009.