Network Analysis of the FlyWire Connectome

Brendan Harrington and Jackson Sutherland

Department of Computer Science, University of Colorado Boulder

CSCI 3352: Biological Networks

Dr. Aaron Clauset

Dec 16, 2024

Table of Contents

Abstract

The neural connectome of the *Drosophila melanogaster* offers a unique opportunity to explore network science concepts in the context of neuroscience. Using the FlyWire Connectome, a dataset of the full adult female brain (FAFB), this project applies a range of concepts covered in lecture to analyze and interpret its structure and function. Techniques such as stochastic modeling, motif analysis, and other methods to compute general network metrics are employed to gain insight into the organization and dynamics of the connectome. Through integrating these approaches, we aim to provide a comprehensive documentation of the exploratory study of this neural network, demonstrating the applicability of concepts from graph theory in the analysis of complex biological systems. This work not only highlights the versatility of these methods but also underscores the potential for interdisciplinary approaches in advancing our understanding of neural architecture.

Introduction

Drosophila melanogaster, commonly known as the fruit fly, has served as a model organism within the sphere of biological science for well over a century (Bellen, 2010). Since their “official” debut as the red-eyed or white-eyed subjects of the classical study on genetic inheritance (Morgan, 1910), adult *Drosophila* have been the topic of countless other studies extending beyond the scope of genetics, with some of the most notable impacts being towards the field of neuroscience. Despite its small size, an adult *Drosophila* is capable of exhibiting complex behavioral patterns such as exceptional motor control and navigation while walking/flying, cognitively-involved decision making, and other socially-motivated performances (Sokolowski, 2001). Additionally, due to their rapid generation time and low cost of production, *Drosophila* specimens are highly accessible to researchers across the globe.

As such, adult fruit flies have become an indispensable asset for studying the neural mechanisms associated with cognition and behavioral phenomena. Up until recently, the only full-brain connectome available has been for an adult *Caenorhabditis elegans*, which has been utilized by the scientific community to perform analyses and simulations on neural circuitry (Cook et al., 2019). Subsequent developments in the connectomics field include the publishing of full mappings of the central brain of an adult *Drosophila*, which proved to be helpful but ultimately incomplete (Scheffer, 2020).

Recently, annotated and proofread data for the full-brain network of an adult *Drosophila* was made available through the FlyWire Codex (https://codex.flywire.ai). Additionally, the original network analysis results were published by Dorkenwald et al. (2024), enabling verification and refinement of methodologies documented in the literature. This project leverages the FlyWire dataset to explore and expand upon these analyses, incorporating methods and concepts discussed throughout the semester. This work highlights the intersection of neuroscience and computational modeling, contributing to a deeper understanding of neural network architecture and its relationship to behavior.

Methods

Dataset

The FlyWire Codex includes an extensive list of datasets with varying file formats, each capturing a different aspect of the *Drosophila* connectome. Contrary to other datasets on Codex with 3-dimensional coordinate info or detailed metadata on graph components, the thresholded “connections” dataset was implemented in this project. Out of the dataset selection on Codex, there are several connection datasets similar to the edge

lists discussed in lecture. The five columns of the datasets are as follows:

- pre_root_id
- post_root_id
- neuropil
- syn_count
- nt_type.

The first two columns represent the source and target of the directed edge, with the remaining 3 columns describing the region of the brain the edge belongs to (neuropil), the synapse count of the edge (syn_count), and the predicted type of neurotransmitter involved in the neural interaction (nt_type).

This format is consistent among the different connection datasets, with the primary difference between the datasets being the filtration of edges. The size of the dataset was cut down by means of filtering out edges with a synapse count of less than 5, which results in a more manageable network that still appropriately represents the original connectome. This threshold also filters out any potential spurious connections and improves our confidence in the accuracy of the data, as it is possible the ML model used to segment the data misidentified low synapse connections.

Software and Computational Environment

To analyze the connectome of *Drosophila melanogaster*, all computational efforts were made using Python and its software libraries. General-purpose utility packages such as matplotlib and numpy were used for data visualization and numerical computations. For network-specific analyses, packages like NetworkX and graph-tool were instrumental in handling the dataset, enabling efficient exploration and manipulation of large-scale neural connectivity data.

The analyses were performed using an assortment of cloud-based and local operating systems to accommodate the varying computational demands of different methods. For computationally intensive tasks, Google Colab and an Oracle Cloud Compute instance were used for their high processing capabilities

and RAM allocation. These cloud platforms facilitated efficient execution of resource-heavy analyses and reduced processing time.

For lighter computational tasks, local UNIX-based machines were utilized due to graph-tool's incompatibility with other operating systems. These systems required the installation of all relevant software and package dependencies, ensuring compatibility and functionality for seamless transitions between local and cloud-based environments. Remote repositories on GitHub were maintained throughout the duration of the project to simplify the process of getting scripts between machines, while ensuring that all copies of the code were consistent and updated fully. This hybrid approach allowed for optimal resource utilization and flexibility throughout the analysis process.

graph-tool

For large-scale graph processing and analysis, the Python package graph-tool has been seen to be orders of magnitude faster than alternatives such as NetworkX (Peixoto, 2020). While many libraries provide tools for graph creation, manipulation, and analysis, graph-tool distinguishes itself through its C++ backend and its efficient handling of graph/edge/vertex properties. These differences make it particularly well-suited for large datasets, such as the FlyWire connectome.

A defining feature of graph-tool is the stricter convention that requires properties to be explicitly defined as PropertyMaps for an associated graph. These maps are tightly integrated with the graph's data structure, improving memory efficiency and enabling vectorized operations directly on the graph. Data for the property maps of a graph can be accessed and adjusted as needed, with a variety of options available for "quick iteration" and other optimization methodologies

The FlyWire connectome is constructed as a directed graph in graph-tool, with neurons as nodes and synaptic connections as edges. Property maps store biological data, such as

synaptic counts (syn_count) and neurotransmitter types (nt_type) for edges, and root IDs and classifications for nodes. This enables efficient querying, filtering, and aggregation of data.

Property maps simplify subgraph generation, such as filtering by neurotransmitter type or implementing higher synaptic thresholds, allowing rapid data transformations. For community detection, graph-tool uses the stochastic block model (SBM) to identify hierarchical modules efficiently, revealing functional clusters in the dataset. These features make the graph-tool workflow efficient for analyzing and interpreting large-scale neural networks like the FlyWire connectome.

Network Analysis

The FlyWire dataset varies from the other networks discussed in class, as there is significantly more emphasis put on the edge interactions occurring within the network. While there are typical conventions present such as edge weights and directed-ness of edges, there is additional metadata associated with the elements of the graph that allow for a deeper analysis. As such, “parallel” edges between two nodes exist within the dataset, which must be properly accounted for during any processing of the data. By understanding the logistics and format of the dataset and all relevant software and package libraries, the neural network of *Drosophila melanogaster* can be efficiently analyzed.

To address the fundamental concept in graph theory, the nodes of the graph are defined to represent neurons and the edges are defined to represent synaptic connections between neurons. The weight of each edge represents its synapse count and hence “connection strength”. The other additional metadata associated with the dataset that was applied within the scope of this work was the neuropil tag of the synaptic interaction (edge) in the graph. By evaluating this aspect of the dataset, the physical sectioning along with the inter- and intra-communication within an adult

Drosophila brain can be explored, along with any other statistics unique to that brain region specifically. An additional metric classifying the type of predicted neurotransmitter associated with the interaction was provided, but the applications of this factor of the connectome were out of the scope of this project.

The general interconnectivity of the network was evaluated by the detection of weakly-connected components (WCCs) and strongly-connected components (SCCs). By treating the graph as directed (SCC) and undirected (WCC), the reachability of other nodes in the graph from an arbitrary start node is iteratively evaluated. The maximum size component in relation to the graph as a whole was selected as the defining result for this aspect of the project, as it provides a good representation of the overall flow of information within the network.

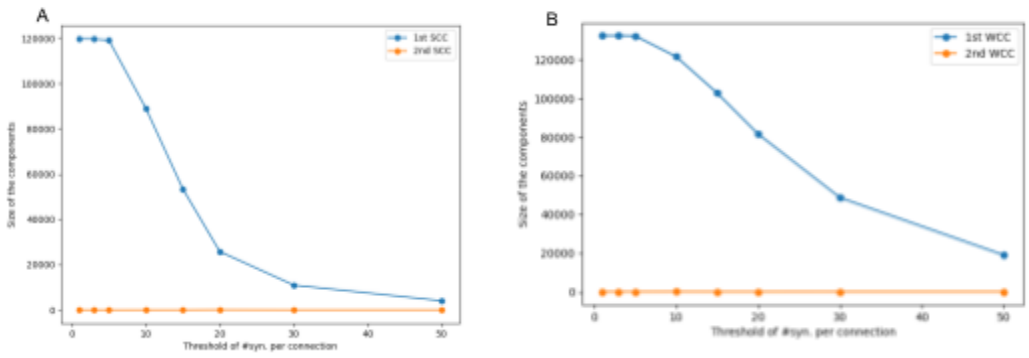
Another convention adhered to throughout the project was the idea of in-/out-degree of nodes, represented as the total number of edges incident with them. By using this perspective rather than negative and positive edge contributions, the general connectedness of the individual neurons within the network can be highlighted.

Results

Connectivity

To analyse the connectivity of the network, we looked at the largest strongly connected component (SCC) and the largest weakly connected component (WCC). We found that the largest SCC contained 89.2496% of all nodes within the network, while the largest WCC contained 98.7345% of all nodes within the network. We also found that the SCC had a mean geodesic path length (MGL) of 4.4828 and the WCC had a mean geodesic path length of 3.9608. These metrics describe the robustness of the network's connectivity. The large percentage of neurons present in the SCC indicates that the network is highly interconnected, indicating that the majority of neurons are able to participate in bidirectional communication. Similarly, when we look at the largest WCC, we see that almost all the neurons are able to participate in one-way communication. The short MGL of both the SCC and WCC indicate that information is able to flow efficiently across most of the network.

Fig. 1: Connectivity



The encapsulation of the largest SCC suggests that the network is resilient to disruptions and can continue to function when edges and nodes are removed. This is shown in Fig. 1, where the connected nature of the network persists even after pruning of edges with a synapse count less than 10, which is around 70% of the network (2,877,573 edges).

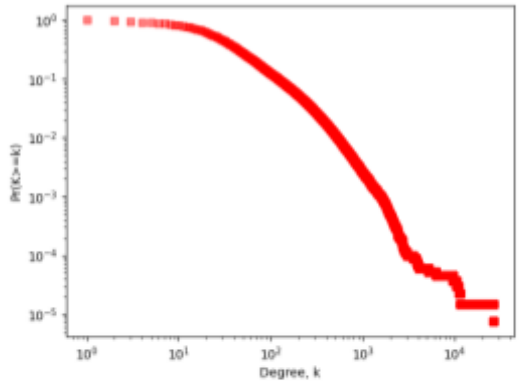
Network Statistics

While not explicitly calculated and mentioned by the authors of the FlyWire literature, the complementary cumulative distribution function (CCDF) can be a good visual representation of the degree distribution of an otherwise extremely heavy-tailed distribution of values. As seen in Fig. 2, the CCDF smooths out the jagged components present in histogram or semi-log representations of degree distributions. In the weighted and unweighted representations of the CCDF, the plots start off with a fairly flat rate of change, which begins decreasing for both representations around a degree of 100. This can be interpreted as the "threshold" at which the count of nodes with a degree that large starts to decrease significantly, which would make sense considering the calculated average in-/out-degree for the nodes in the graph was found to be about 254.

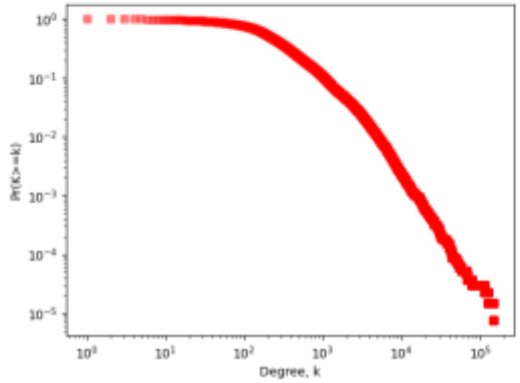
Additional calculations for more general graph metrics such as reciprocity and clustering

Fig. 2 CCDF

a. The unweighted CCDF.



b. The weighted CCDF.



coefficient were also similar to the results mentioned in the literature. The reciprocity of 0.147 obtained in this study is comparable to the reciprocity of 1.38 obtained in the original study, with the increase in reciprocity most likely due to the increase of low-degree but connected nodes in the v783 snapshot of the dataset. In this study, the clustering coefficient was found to be 0.0487 with a standard

deviation of 0.01355, which lines up well with the result of 0.0477 from the original paper. These measurements were obtained using methods discussed in class, but were optimized and wrapped in convenient functions within the graph-tool library.

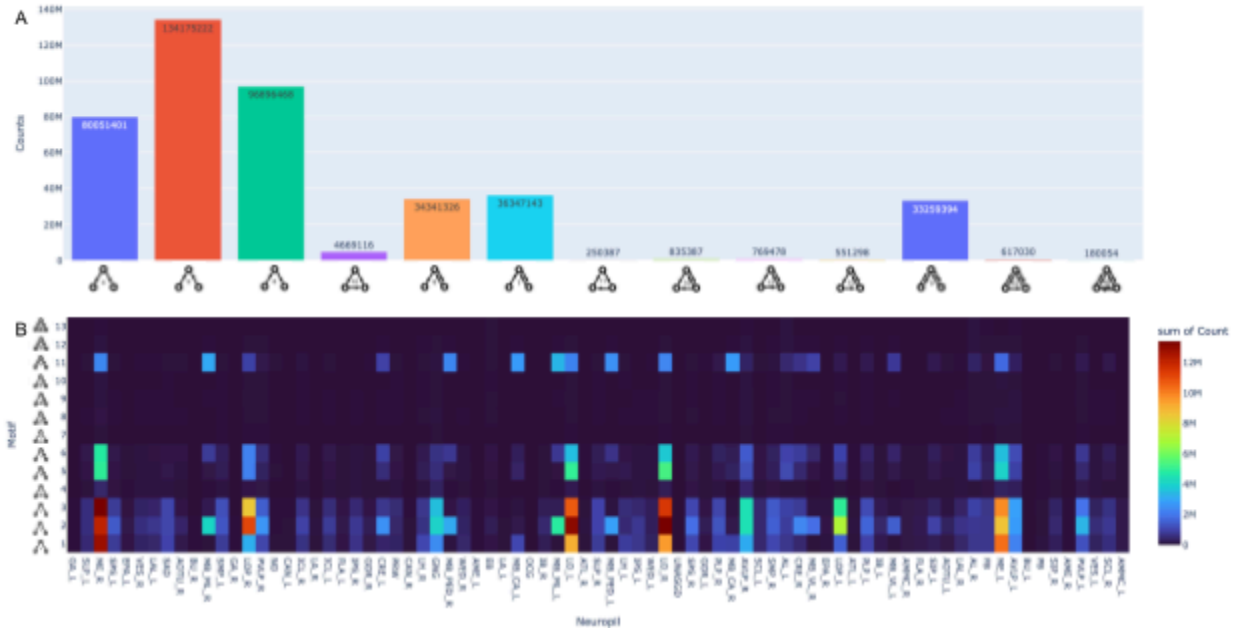
Supplementary plots with additional data about the general graph statistics can be found in the Appendix, providing a basis for more thorough analysis and comparison of the degree distributions. By making use of the graphs provided in the literature by Dorkenwald et al., the data analysis procedure was streamlined and the computational needs of the project's workflow were reduced greatly. These baseline graphs were assumed to be accurate, with our calculations representing the quality of the original data.

Motifs

To understand the underlying structural patterns within the network, we performed motif analysis to detect subgraphs of size three, commonly referred to as triads. Some of these motifs have known functional roles in the network, whereas others are more speculated about. The presence of these motifs gives us insight into the overall structure of the network and how information is locally processed as it flows through the network. To achieve this, we utilized the graph-tool library, and analyzed the network as a whole, as well as individual neuropils.

For the entire network we found 422,943,704 total motifs, with the 2-sink motif representing 31.72% of the total motifs in the network (Fig 3a). To explore the variation within brain regions, we filtered the graph by neuropil and ran motif detection on each subgraph generated. This provides us with a more localized view of motif distribution and insight into the functional roles of different brain regions.

Fig. 3: Motifs



We found a high density of motifs in the medulla, lobula and lobula plate, all located within the optic lobe (Fig 3b). There are 5 specific motifs found in this region, potentially reflecting their role in processing and integrating visual information. The high density of motifs found in these areas indicate highly interconnected local circuits, which may play a role in efficient signal processing and support functions such as motion detection, spatial awareness, and visual memory formation.

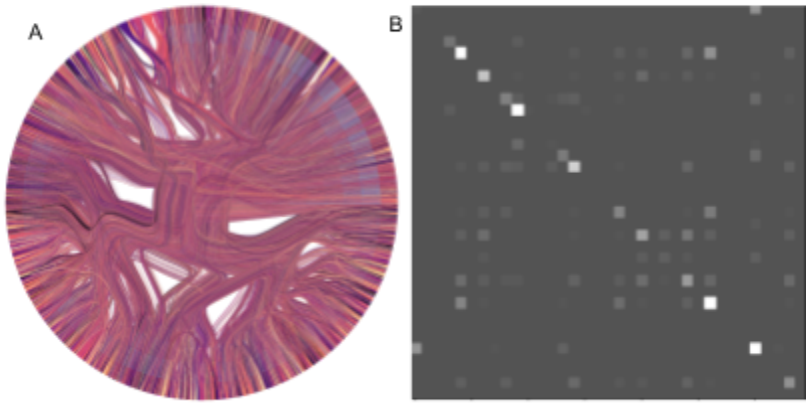
By examining motif distribution among neuropils, it is possible to gain insight into the functional roles that specific brain regions play and how their connectivity patterns facilitate these functions. The localized motif distribution observed within the optic lobe provides strong evidence for its specialization of visual information processing and integration.

Community Detection

Community detection is a key component in identifying high density regions of networks. These clusters can often correspond to functional or structural modules in real world networks, such as anatomical or functional regions in the brain network.

In this paper, we conducted preliminary analysis of community detection using graph-tools built in stochastic block model and nested stochastic block model. While this analysis was limited, it revealed some interesting connection patterns that warrant further exploration in future work.

Fig. 4: Community Detection



As shown in Fig. 4b, there are multiple regions of high reciprocity. These could represent areas of integrated information processing, where information can easily flow bidirectional. These reciprocal patterns suggest that there are strong feedback mechanisms and may facilitate local information processing. These areas may correspond to functions such as sensory integration, motor control, or other higher-order cognitive tasks.

Discussion

Key Findings

Much like the results from the original work by Dorkenwald et al., this project found a strong connectivity within the connectome of Drosophila melanogaster. While procedures as detailed as those in the literature were out of the scope of the semester-long project, the basic concepts were applied during the process, capable of being expanded upon as needed. Our results confirm the robustness of the network, and highlight the key role that motifs play in information processing within certain regions of the brain.

One of the key findings was the robustness of the network, shown by the size and connectivity of the largest strongly connected component (SCC) and weakly connected component (WCC). The largest SCC contained 89.25% of all nodes, while the largest WCC contained 98.73% of nodes, exemplifying the network's resilience to fragmentation and capacity for information propagation.

Motif analysis provided important insight into the functional organization of the brain. We identified over 422 million motifs in the full network, with the 2-sink motif being the most prevalent, accounting for 31.72% of all detected motifs. This high prevalence of specific motifs could point to their role in localized information processing and feedback mechanisms, particularly in regions such as the medulla, lobula, and lobula plate within the optic lobe. These regions exhibited a high density of motifs, reflecting their specialization in visual information processing and integration.

Preliminary community detection using the stochastic block model and nested stochastic block model revealed interesting patterns of modularity. Multiple regions of high reciprocity were identified, suggesting tightly integrated clusters, possibly corresponding to feedback loops or local processing. These findings, while limited in scope, provide a good

starting point for future work on the connectome's functionality.

Overall, our results reinforce the previous work on the Drosophila connectome, showing a highly organized and robust network, with motifs and community structures playing critical roles in its capacity for information processing.

Challenges and Limitations

With a dataset as large as the neural mapping of *Drosophila*, measures needed to be taken to maximize productivity throughout the course of the semester. The unprocessed network data is contained within over 22 million rows (edges) worth of data. With this being the case, a degree of precision and accuracy was sacrificed to reduce the amount of data that needed to be processed.

Rather than manipulating the mass amount of data that composed the full connectome, the thresholded dataset was selected, as its 3.8 million rows seemed much more realistically manageable. This choice was deemed appropriate, as the lowest weight edges do not contribute a huge amount to the overall structure of the graph. Additionally, the more updated metadata included in the non-thresholded datasets was less applicable to this project, furthering the justification for selecting a more filtered graph.

Another limitation was the dataset we are using is still not fully complete. While every neuron in the *Drosophila melanogaster* is accounted for, there are specific types of neurons, called gap-junction neurons, that are not included in the data set. These neurons are speculated to play an important role in the brain during developmental phases. While the function of these neurons in an adult brain is still unknown, including them in the dataset could highlight the functional roles they play in the adult brain as well as give us a broader understanding of the network dynamics.

Future Work

One aspect of the project that stood out was the motif analysis within the optic lobe. The high concentration in this region warrants more exploration and detailed analysis, incorporating the role that neurotransmitters play and deepening our understanding of how these motifs affect the visual signals being processed.

Another key area is community detection. While our analysis revealed regions of high reciprocity, comparing these results to null models could provide deeper insights into the organization of the connectome. By examining how communities interact with one another and comparing these structures with known anatomical and functional regions, future work could expose the principles underlying how the brain is able to efficiently process information.

Additionally, expanding the analysis to include the non-thresholded dataset could provide a more comprehensive view of the connectome. Incorporating lower connection strength synapses could uncover weak but functionally significant connections and expand our understanding of long-range communication within the brain. As of right now though, this could prove to be challenging as it is unknown how accurate the lower strength connections are.

By building on this work, our future findings can help advance our understanding of the *Drosophila melanogaster* connectome even further and illuminate its broader implications for neuroscience as a whole.

References

1. Bellen, H. J., Tong, C., & Tsuda, H. (2010). 100 years of Drosophila research and its impact on vertebrate neuroscience: a history lesson for the future. *Nature Reviews Neuroscience*, *11*(7), 514-522.

2. Morgan, T. H. (1910). Sex limited inheritance in Drosophila. *Science*, *32*(812), 120-122.

3. Sokolowski, M. B. (2001). Drosophila: genetics meets behaviour. *Nature Reviews Genetics*, *2*(11), 879-890.

4. Cook, S. J., Jarrell, T. A., Brittin, C. A., Wang, Y., Bloniarz, A. E., Yakovlev, M. A., ... & Emmons, S. W. (2019). Whole-animal connectomes of both Caenorhabditis elegans sexes. *Nature*, *571*(7763), 63-71.

5. Scheffer, L. K., Xu, C. S., Januszewski, M., Lu, Z., Takemura, S. Y., Hayworth, K. J., ... & Plaza, S. M. (2020). A connectome and analysis of the adult Drosophila central brain. *elife*, *9*, e57443.

6. Dorkenwald, S., McKellar, C. E., Macrina, T., Kemnitz, N., Lee, K., Lu, R., ... & Seung, H. S. (2022). FlyWire: online community for whole-brain connectomics. *Nature methods*, *19*(1), 119-128.

7. Peixoto, T. P. (2020, July 8). *Graph-tool performance comparison*. graph-tool. https://graph-tool.skewed.de/performance.html

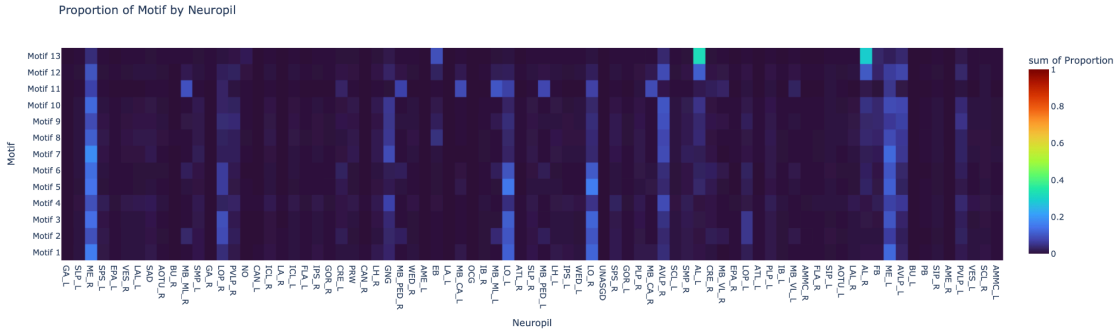
8. Zheng, Z., Lauritzen, J. S., Perlman, E., Robinson, C. G., Nichols, M., Milkie, D., ... & Bock, D. D. (2018). A complete electron microscopy volume of the brain of adult Drosophila melanogaster. *Cell*, *174*(3), 730-743.

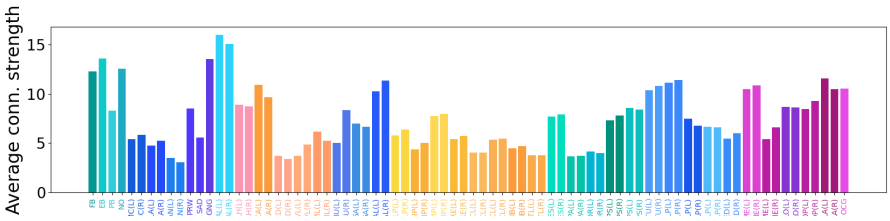
9. Tolwinski N. S. (2017). Introduction: Drosophila-A Model System for Developmental Biology. Journal of developmental biology, 5(3), 9. https://doi.org/10.3390/jdb5030009

10. Dorkenwald, S., McKellar, C. E., Macrina, T., Kemnitz, N., Lee, K., Lu, R., Wu, J., Popovych, S., Mitchell, E., Nehoran, B., Jia, Z., Bae, J. A., Mu, S., Ih, D., Castro, M., Ogedengbe, O., Halageri, A., Kuehner, K., Sterling, A. R., Ashwood, Z., … Seung, H. S. (2022). FlyWire: online community for whole-brain connectomics. Nature methods, 19(1), 119–128. https://doi.org/10.1038/s41592-021-01330-0

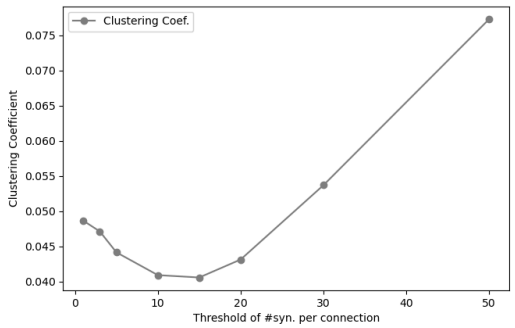
11. Dorkenwald, S., Matsliah, A., Sterling, A. R., Schlegel, P., Yu, S. C., McKellar, C. E., ... & Murthy, M. (2024). Neuronal wiring diagram of an adult brain. Nature, 634(8032), 124-138.

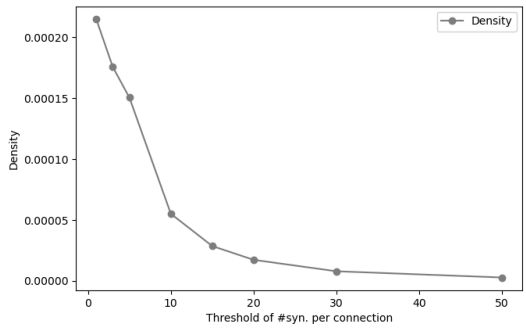
Figures

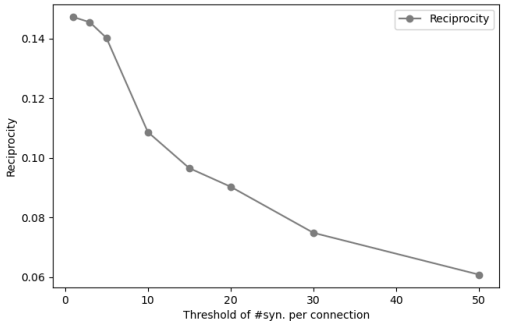


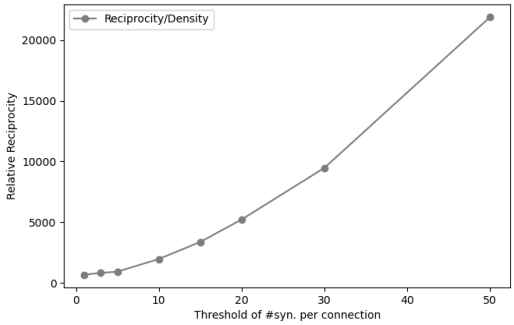


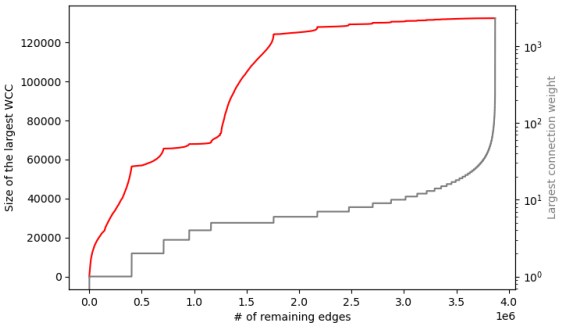

```
top sizes of SCC: [119756      4      4      4      4      4      4      3      3      3]
number of SCC: 14337

top sizes of WCC: [132483     17     16     11     10     10     10     10     10     10]
number of WCC: 244
```

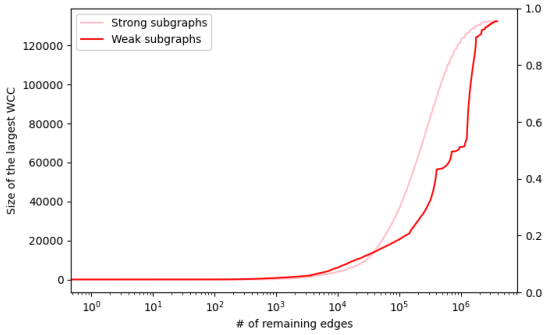


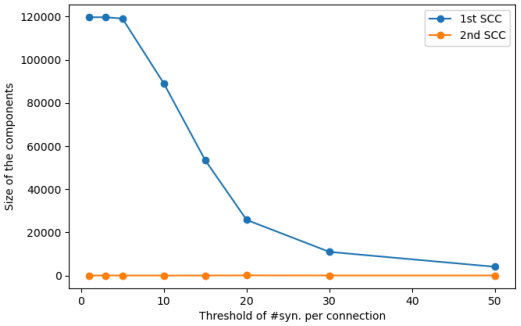


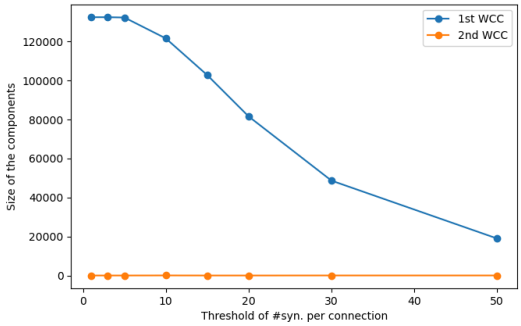

Reciprocity
Reciprocity
Threshold of #syn. per connection

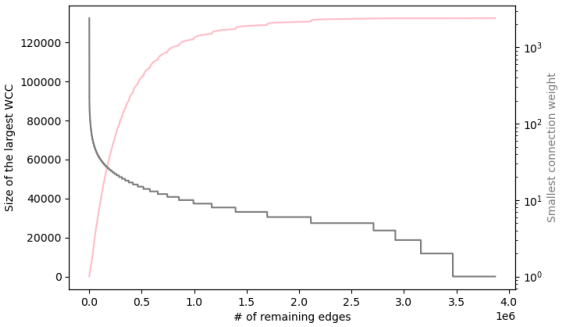

Reciprocity/Density
Relative Reciprocity
Threshold of #syn. per connection


Size of the largest WCC
Largest connection weight
# of remaining edges
1e6

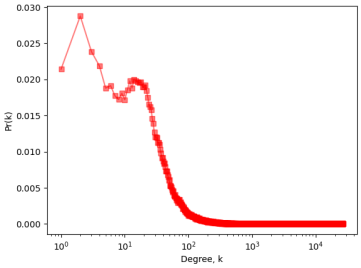

Strong subgraphs
Weak subgraphs
Size of the largest WCC
# of remaining edges

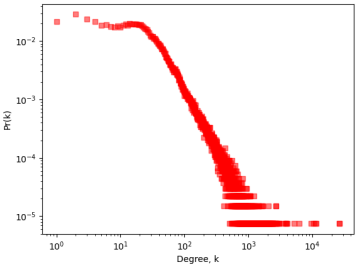

1st SCC
2nd SCC
Size of the components
Threshold of #syn. per connection

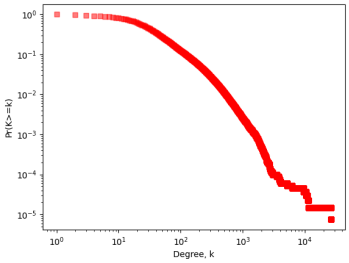

1st WCC
2nd WCC
Size of the components
Threshold of #syn. per connection


Size of the largest WCC
Smallest connection weight
# of remaining edges
1e6

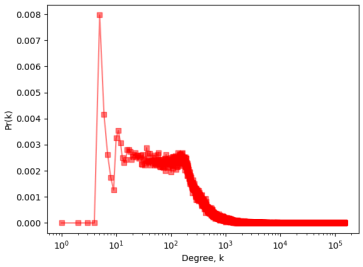
Unweighted

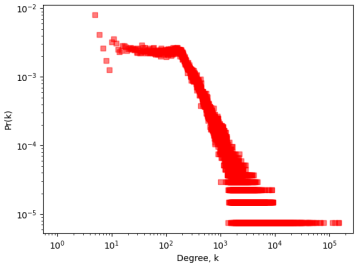






Weighted





Appendices

All code is available at our github repository.

The data used for this analysis is available at flywire.ai