Community detection and classification in hierarchical stochastic blockmodels

Minhcent Tangzynski

Applied Mathematics and Statistics
Johns Hopkins University

Introduction and Overview

1. The problem of deciding whether two given graphs are the "same" arises in neuroscience and social networks.
2. Many large graphs are composed of loosely connected smaller graph primitives.
3. Existing community detection algorithms focus mostly on uncovering the subgraphs themselves. We want to identify, classify, and characterize *stochastically similar* structures in big graphs.
4. We propose a valid and consistent test for testing the hypothesis that two graphs are from the same random graph distribution. The test proceeds by embedding the graphs into a lower-dimensional space and then computing a kernel-based distance between the embedded points.

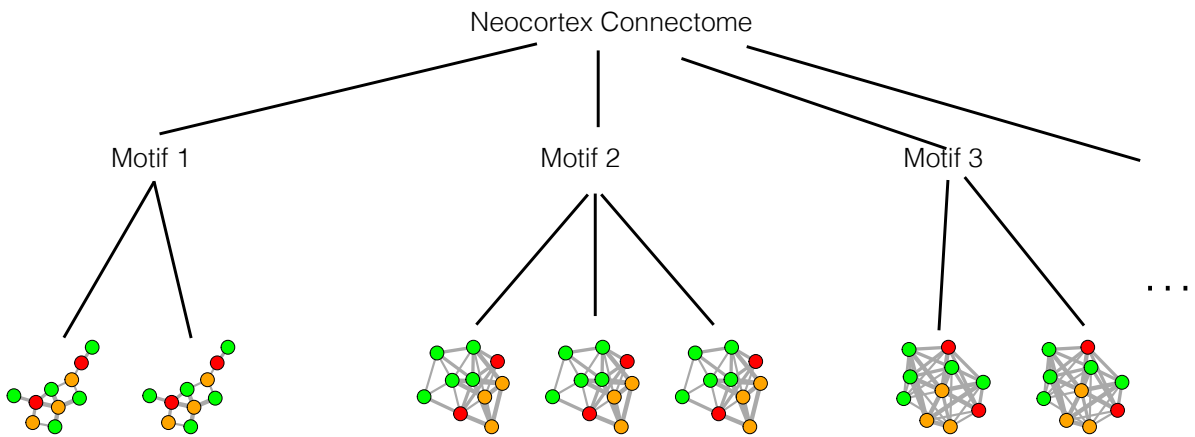
Outline

The cortical column conjecture

The cortical column conjecture stipulates that the neocortex is a graph of neuronal connections that exhibits motifs representing repeated processing modules.

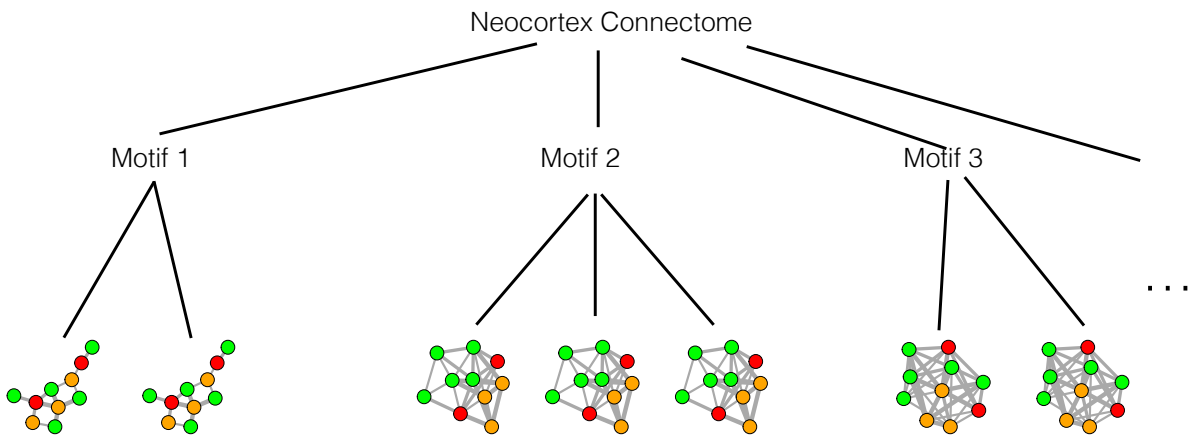
The cortical column conjecture

The cortical column conjecture stipulates that the neocortex is a graph of neuronal connections that exhibits motifs representing repeated processing modules.



The cortical column conjecture

The cortical column conjecture stipulates that the neocortex is a graph of neuronal connections that exhibits motifs representing repeated processing modules.



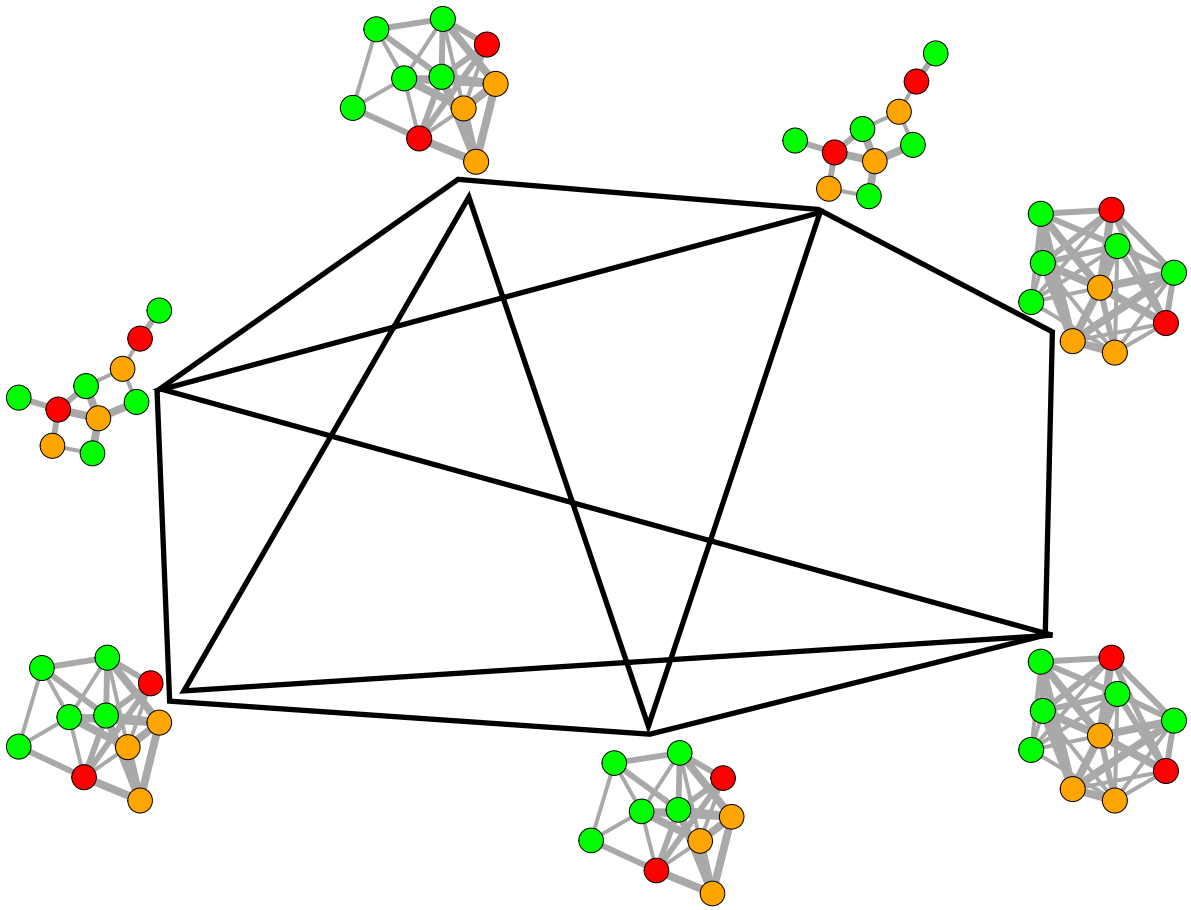
We interpret *motif* as a collection of *stochastically similar* subgraphs.

Testing the cortical column conjecture

- Given the full neocortex connectome, how would we go about testing the cortical column conjecture?

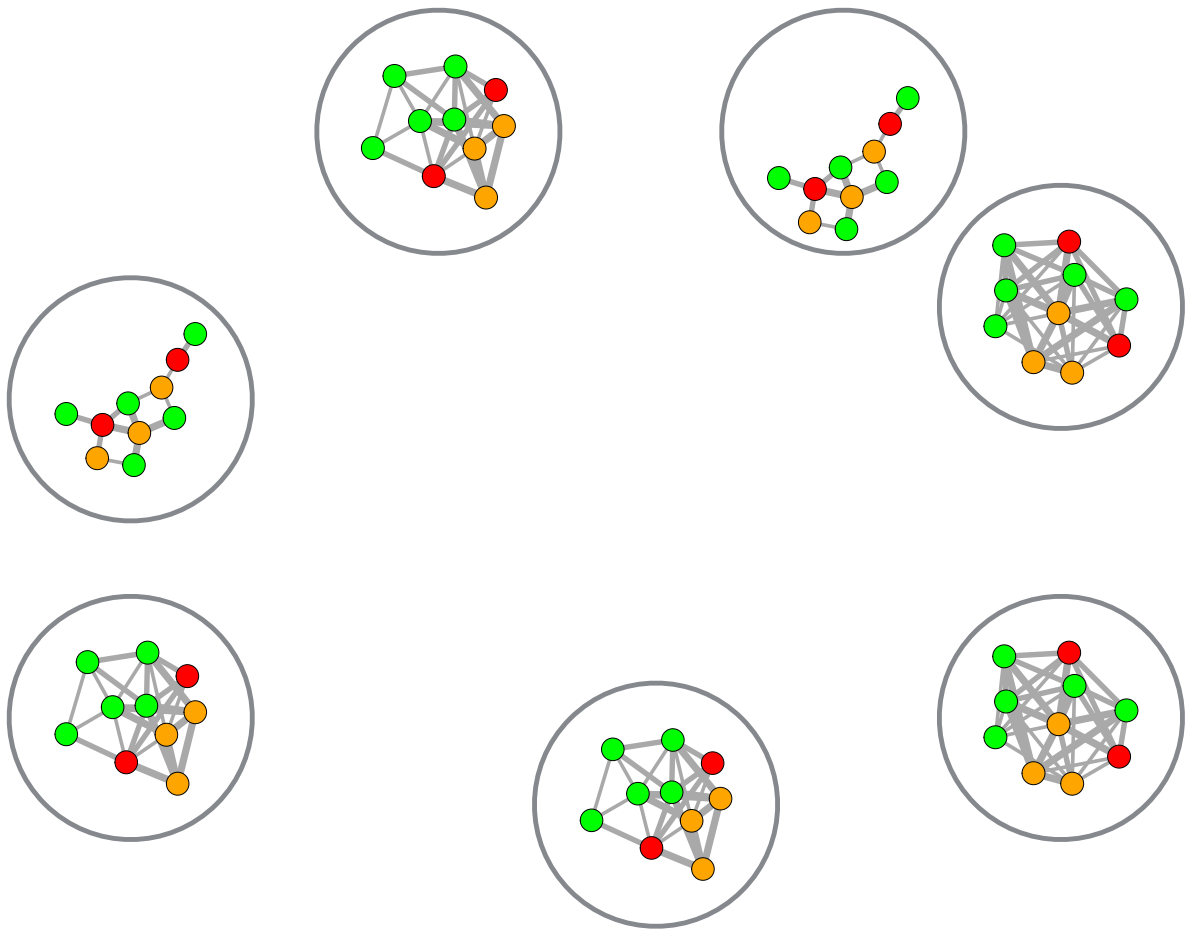
Testing the cortical column conjecture

- Given the full neocortex connectome, how would we go about testing the cortical column conjecture?
- We would need to simultaneously find the processing modules (community detection) and classify them into motifs (classification)



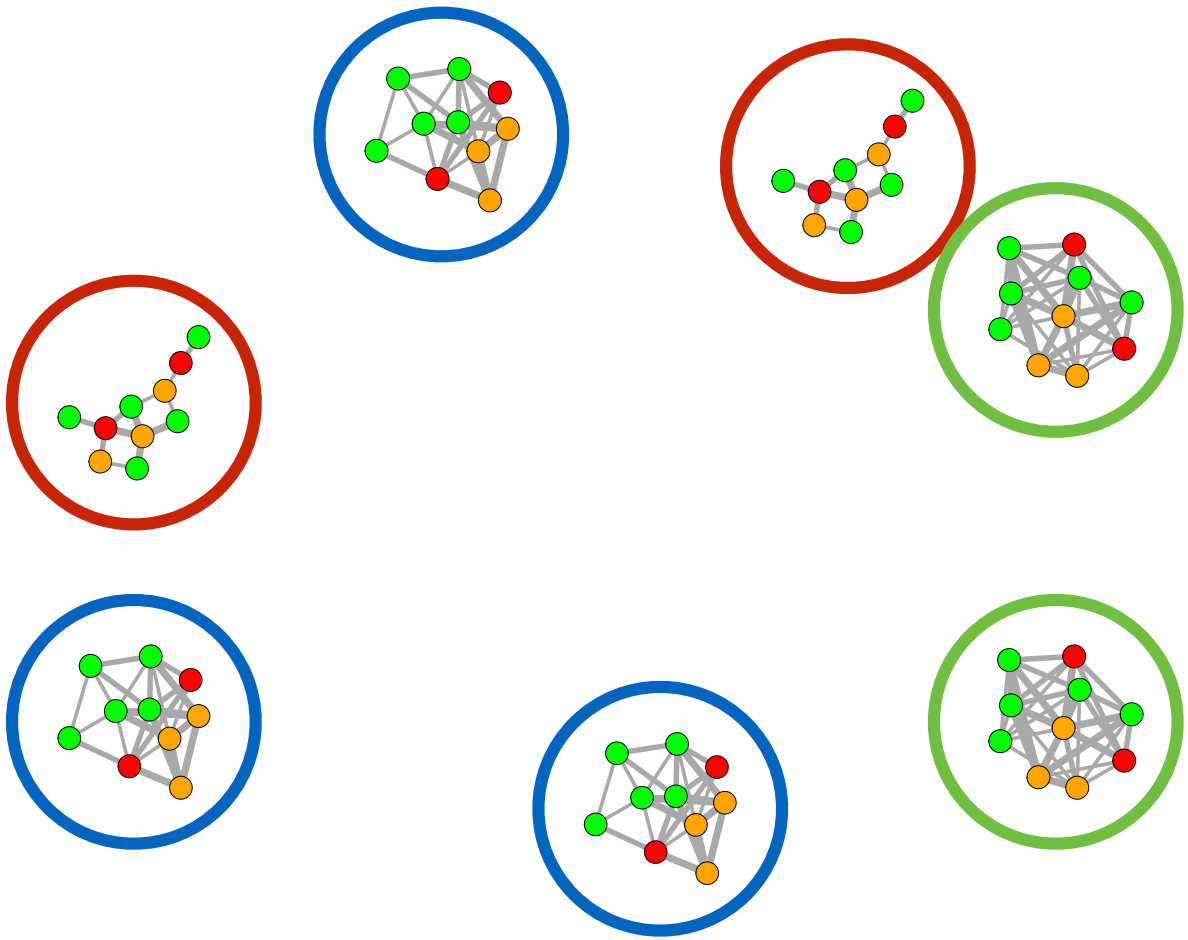
Testing the cortical column conjecture

- Given the full neocortex connectome, how would we go about testing the cortical column conjecture?
- We would need to simultaneously find the processing modules (community detection) and classify them into motifs (classification)



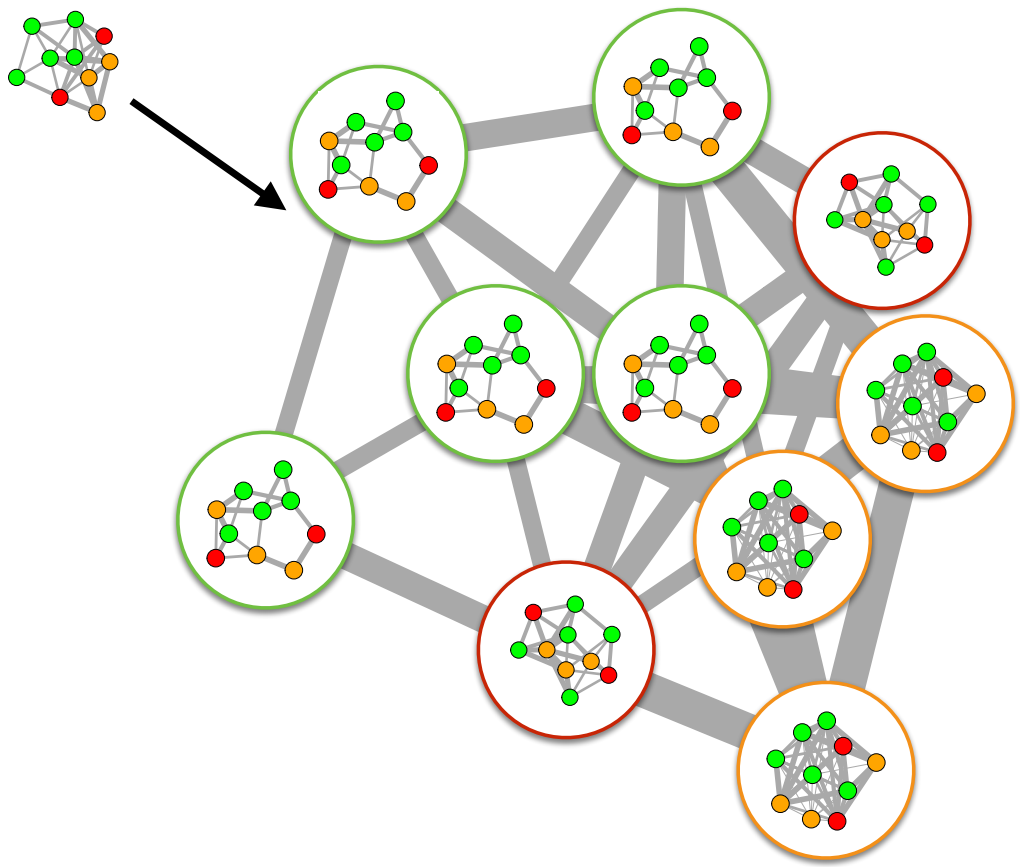
Testing the cortical column conjecture

- Given the full neocortex connectome, how would we go about testing the cortical column conjecture?
- We would need to simultaneously find the processing modules (community detection) and classify them into motifs (classification)



Testing the cortical column conjecture

- Given the full neocortex connectome, how would we go about testing the cortical column conjecture?
- We would need to simultaneously find the processing modules (community detection) and classify them into motifs (classification)–and then recurse on each motif



What does the data look like ?

- On a small scale, high resolution data consists of only a few hundred partial neurons.
- However, massively-funded effort are underway to collect $1\mathrm{mm}^3$ mammalian cortex via electron microscopy.

  It is commonly believed that neural algorithms employ data representations, transformations, and learning rules that are conserved across stages. It should therefore be possible to apprehend the neural computations underlying information processing ... by interrogating a small fraction of the entire cortex, so long as that fraction is judiciously selected ...

  – From the MICrONS BAA

Outline

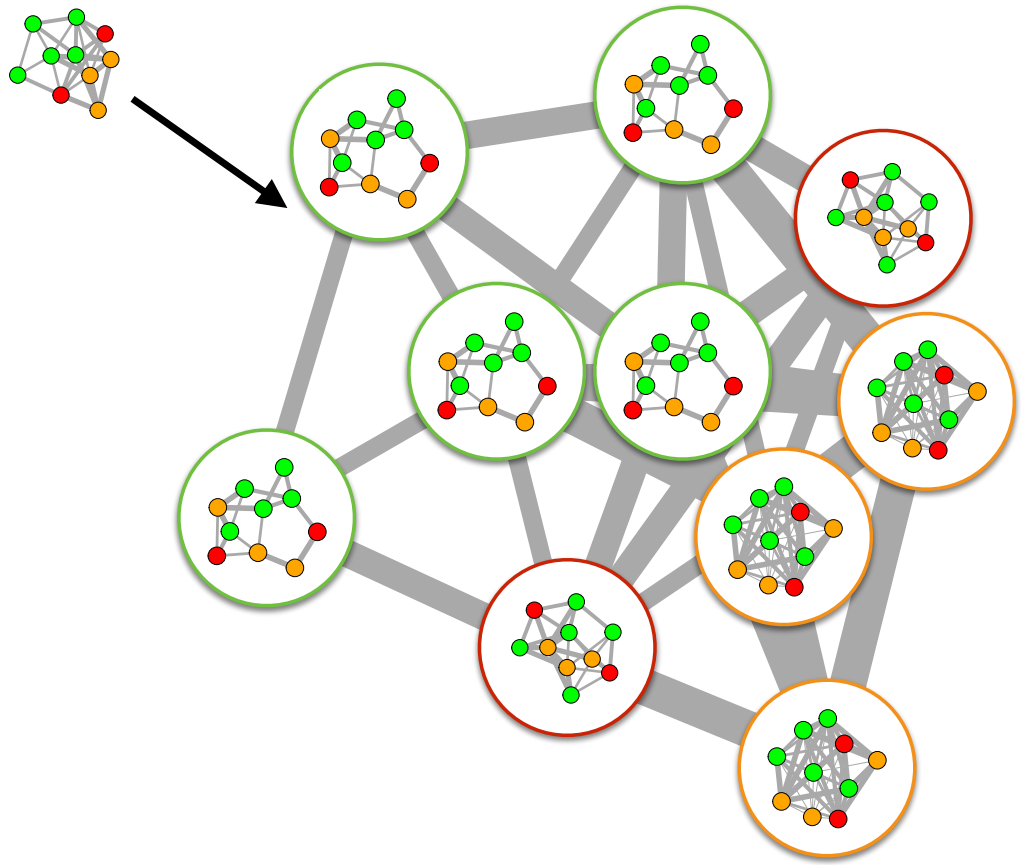
Random dot product graphs (RDPGs)

Random dot product graphs (RDPGs) are an example of *latent position graphs*:

- each vertex is associated with a latent position (a vector)
- edges are conditionally independent
- probability of an edge between two vertices i and j, $\mathrm{p_{ij}}$, is a function (here, the dot product) of the latent positions.
- Stochastic blockmodels are graphs where the probability $\mathrm{p_{ij}}$ of an edge between two vertices i and j depends only on the *block memberships* of the vertices, and these can also be regarded as RDPGs

Hierarchical stochastic blockmodels

In hierarchical stochastic blockmodels (HSBMs), there are subgraphs that are loosely connected between and more densely connected within; each of these subgraphs is itself another stochastic block model.



HSBM (an example)

For simplicity, we present a sample adjacency matrix of an HSBM, where the θ_i's characterize the stochastic block model subgraphs. Two induced subgraphs H_r and H_s are said to be from the same motif if, for example, $\theta_r = \theta_s$.

$$\begin{pmatrix} \left[\mathrm{SBM}(\theta_1)\right] & \left[\mathrm{ER}(p)\right] & \cdots & \left[\mathrm{ER}(p)\right] \\ \left[\mathrm{ER}(p)\right] & \left[\mathrm{SBM}(\theta_2)\right] & \cdots & \left[\mathrm{ER}(p)\right] \\ \vdots & \vdots & \ddots & \vdots \\ \left[\mathrm{ER}(p)\right] & \left[\mathrm{ER}(p)\right] & \cdots & \left[\mathrm{SBM}(\theta_Q)\right] \end{pmatrix}$$

A nonparametric two-sample testing problem

Given two adjacency matrices $\mathbb{A}$ and $\mathbb{B}$ for a pair of random dot product graphs, we have developed a *nonparametric test of hypothesis* to determine whether the two graphs are probabilistically "similar" or "different."

This test depends on a convenient representation, called the adjacency spectral embedding, of the adjacency matrix of the graph.

This test works even with graphs of different sizes, and even for graphs where no correspondence between vertices is known.

Motif detection algorithm

Input: Adjacency matrix $A \in \{0,1\}^{n \times n}$
Output: Tree of subgraphs and characterization of their dissimilarity
while Cluster size exceeds threshold **do**
Step 1: Compute the adjacency spectral embedding of A.
Step 2: Cluster the rows of the adjacency spectral embedding to obtain subgraphs $\hat{H}_1, \cdots, \hat{H}_Q$.
Step 3: For each $i \in [Q]$, compute the adjacency spectral embedding for each subgraph $\hat{H}_i$;
Step 4: Compute the nonparametric test statistic for each pair of subgraphs, producing a pairwise dissimilarity matrix on induced subgraphs;
Step 5: Cluster induced subgraphs into motifs according to similarity;
Step 6: Recurse on each motif;
end while

Outline

Motif detection in the Drosophila connectome

- Partial *Drosophilia* medulla connectome (**?**)
  1. First constructed the full connectome between 379 named neurons (believed to be a single column)
  2. Sparsely reconstructing the connectome between and within surrounding columns via a semi-automated procedure.
- 1748 vertices in its largest connected component.
- We embed the largest connected component into $\mathbb{R}^{13}$.
- Cluster into $\widehat{Q} = 8$ clusters of sizes $176, 237, 434, 237, 142, 237, 115,$ and 170 vertices.
- Compute matrix of test stastics between subgraphs.
- Cluster the matrix into motifs (using hierarchical clustering).

Motif detection in the Drosophila connectome (continued)

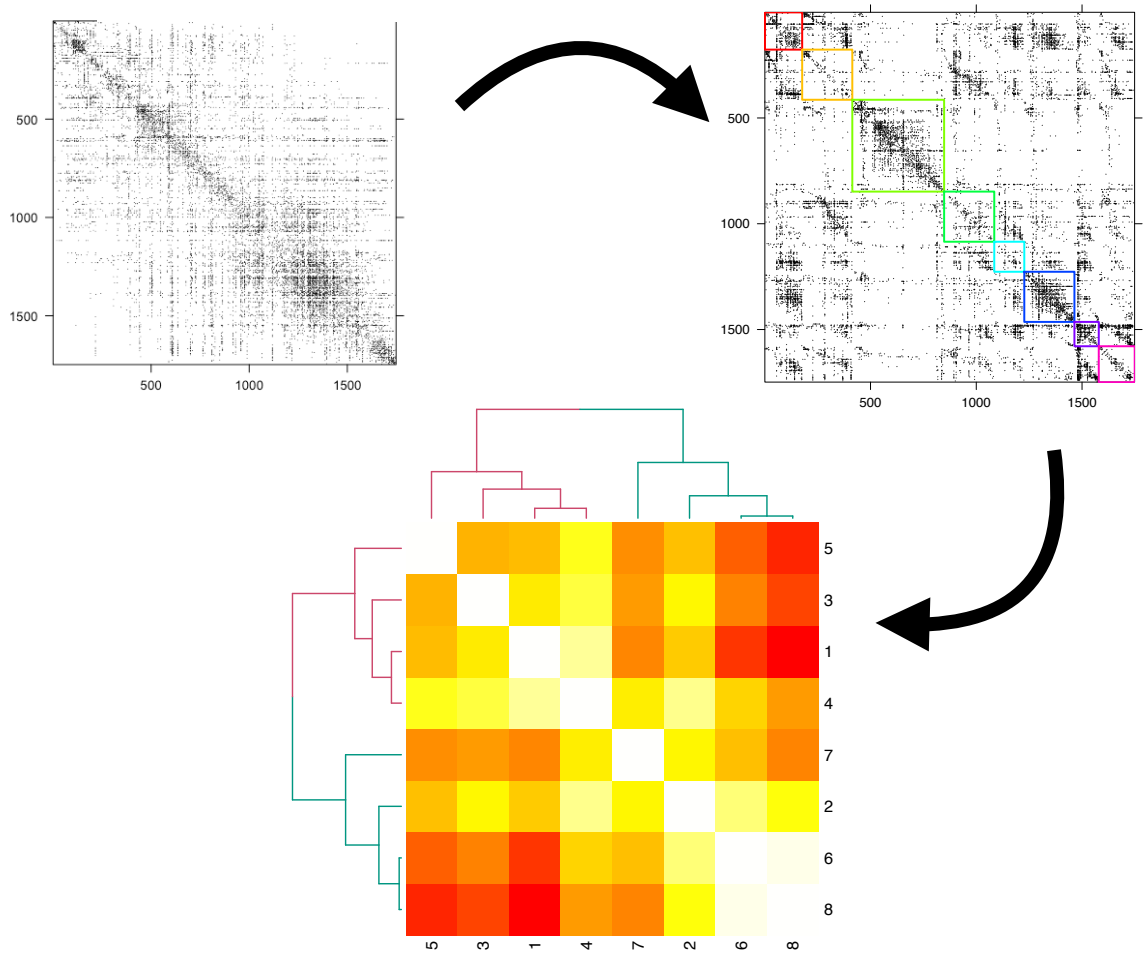


Figure 1: Visualization of our method applied to the *Drosophila* connectome. We show the adjacency matrix (upper left), the clustering derived, and lastly $\widehat{S}$ calculated from these clusters. Clustering the subgraphs based on $\widehat{S}$ suggests two repeated motifs: $\{1, 4\}$ and $\{2, 6, 8\}$. The hierarchical clustering also reveals 2nd level motif repetition within the second motif given by $\{6, 8\}$.

Motif detection in the Friendster network

- The Friendster social network contains roughly 60 million users and 2 billion connections.
- There are roughly 1 million communities at the local scale.
- We expect the social interactions in these communities to inform the function of the different communities; hence we expect to observe distributional repetition among the graphs associated with these communities.
- We embed the Friendster graph into $\mathbb{R}^{14}$ using `FlashGraph`.
- The best coarse-grained clustering of the graph is achieved with $\widehat{\mathrm{Q}} = 15$ large-scale clusters ranging in size from one million to 15.6 million vertices.

Motif detection in the Friendster network (continued)

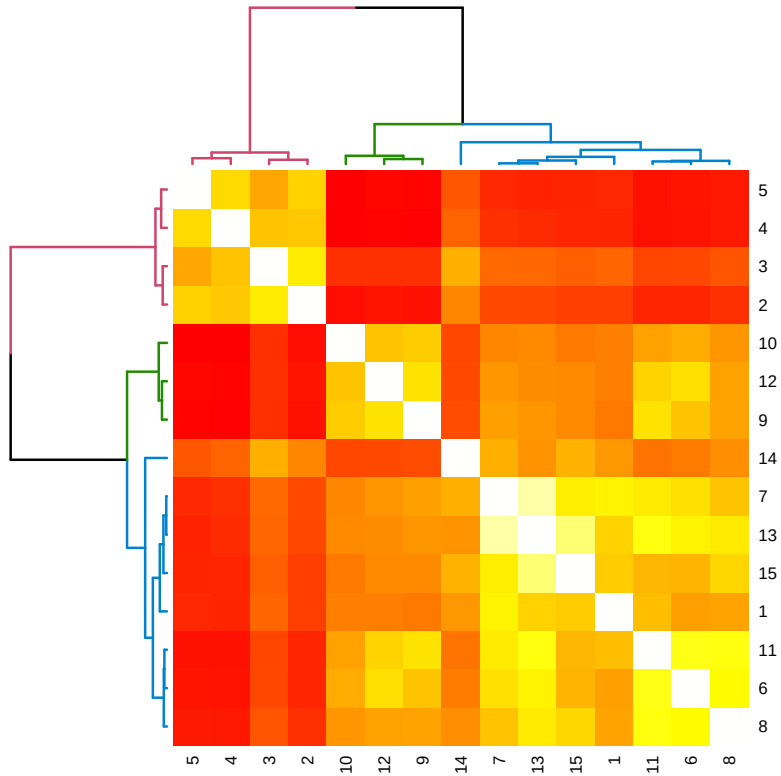


Figure 2: Heat map of similarities among the top-level communities in the Friendster social network. Similarity between any two communities is represented on the spectrum between white and red, with white representing highly similar communities.

Motif detection in the Friendster network (continued)

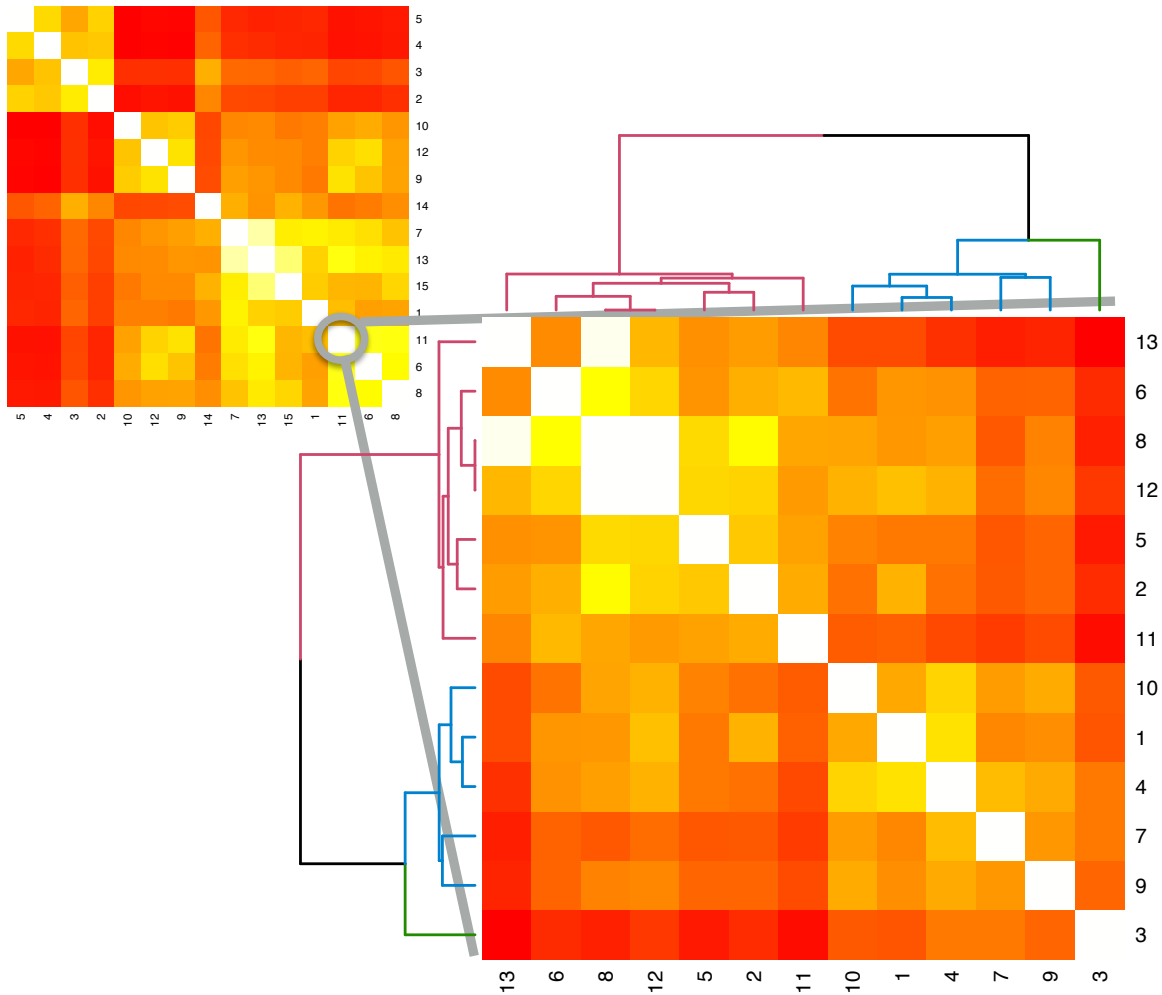


Figure 3: Heat map depiction of the level two Friendster estimated dissimilarity matrix $\widehat{S} \in \mathbb{R}^{13\times 13}$ of $\widehat{H}_{11}$.

We identify similar communities

We identify similar communities at the first level, and then analyze a particular subgraph at the first level to decompose it into similar and dissimilar subgraphs at a subsequent level.

Implications of this for this social network?

At present, unclear–these subgraphs are each very large.

Nevertheless, our approach presents a theoretically-justified approach to community classification, one that is functional even for large networks.

Challenges

Have made quite a few assumptions that make our lives easy.

- Potentially a large number of repeated motifs.
- How to leverage edge and vertex covariates (spatial location, neuron type, . . .).
- Error propagation between steps. In the presence of deeper hierarchy with varying levels of interconnectivity ?
- Leveraging (partial) vertex correspondence or non-existent vertex correspondence (Graph Matching or Parameter Matching) and other network structures.

Thank you!

- M. Tang, A. Athreya, D. L. Sussman, V. Lyzinski, C. E. Priebe. "A non-parametric two-sample hypothesis testing problem for random dot product graphs", Bernoulli, accepted for publication. http://arxiv.org/abs/1409.2344.
- V. Lyzinski, M. Tang, A. Athreya, Y. Park, C. E. Priebe. "Community detection and classification in hierarchical stochastic blockmodels", http://arxiv.org/abs/1503.02115.