# Graphon-Based Information Bottleneck Analysis of Neural Networks via Stochastic Block Models

## Abstract

Deep neural networks are often analyzed through the Information Bottleneck (IB) principle, which formalizes a trade-off between compressing inputs and preserving information relevant for predicting the target variable. Although conceptually appealing, directly estimating mutual information in large architectures is computationally challenging. We propose a graphon-based approach that approximates multilayer perceptrons by fitting weighted stochastic block models (WSBMs) to their weight matrices. The resulting SBM graphons capture the modular structure that emerges during training and enable tractable block-level estimates of $I(X;T)$ and $I(T;Y)$. Our analysis yields preliminary results toward more interpretable IB planes and introduces block-to-block information flow maps, which qualitatively align with classical IB theory. This framework connects graph limit theory with neural network interpretability, offering a scalable geometric abstraction for analyzing information flow in deep models.

## 1. Introduction

The information bottleneck (IB) principle provides a powerful framework for understanding the trade-off between compression and prediction in neural networks (Shwartz-Ziv and Tishby, 2019). Despite its conceptual appeal, applying IB theory to modern architectures is challenging: estimating mutual information at scale requires sophisticated estimators, and often intractable to compute directly.

In parallel, *graphons*—limit objects of sequences of dense graphs (Lovász, 2012)—have emerged as a central tool in network science, supporting the study of contagion, modularity, and information diffusion in large network systems. This perspective suggests a natural extension: using graphons to approximate information flow in deep neural networks.

In this preliminary work, we propose a **graphon representation of multilayer perceptrons (MLPs)** derived from weighted stochastic block models fitted to weight matrices. This abstraction captures the modular structure that emerges during training (Watanabe et al., 2017), yields tractable block-level estimates of mutual information, and provides an alternative analytic route to studying training dynamics.

## 2. Methodology

We approximate the information flow in MLPs by fitting *Weighted Stochastic Block Models (WSBMs)* to layer weight matrices (Ng and Murphy, 2021), and representing the result as an SBM graphon. This enables block-level mutual information estimates that are tractable and interpretable.

### 2.1. Stochastic Block Model Fitting

Given a weight matrix $W \in \mathbb{R}^{n_{\text{in}} \times n_{\text{out}}}$ between two layers, we view it as the adjacency matrix of a bipartite weighted graph: left vertices represent input neurons, right vertices represent output neurons, and edges carry weights $w_{ij}$. We partition input neurons into

$K_{\text{in}}$ clusters and output neurons into $K_{\text{out}}$ clusters. The block mean weight from cluster $a$ to $b$ is

$$\mu_{ab} = \frac{1}{|\{i : z_{\text{in}}(i) = a\}| \cdot |\{j : z_{\text{out}}(j) = b\}|} \sum_{\substack{i: z_{\text{in}}(i)=a \\ j: z_{\text{out}}(j)=b}} w_{ij}.$$

### 2.2. Graphon Representation

Let $\pi_L(a)$ and $\pi_R(b)$ denote the relative sizes of clusters. We define a step-function graphon

$$W(u,v) = \mu_{ab}, \quad \text{if } u \in I_a,\ v \in J_b,$$

where $\{I_a\}, \{J_b\}$ partition $[0,1]$ according to $\pi_L$ and $\pi_R$.

### 2.3. Mutual Information Estimates

We define the joint distribution

$$P(a,b) = \frac{\pi_L(a)\,\pi_R(b)\,|\mu_{ab}|}{\sum_{a',b'} \pi_L(a')\pi_R(b')|\mu_{a'b'}|}.$$

With marginals $P(a) = \sum_b P(a,b)$ and $P(b) = \sum_a P(a,b)$, the block-level mutual information is

$$I(X;T) = \sum_{a=1}^{K_{\text{in}}} \sum_{b=1}^{K_{\text{out}}} P(a,b)\,\log \frac{P(a,b)}{P(a)P(b)}.$$

By assigning each sample to the output cluster with highest mean activation, we build a contingency table $P(c,y)$ between clusters $c$ and labels $y$, yielding

$$I(T;Y) = \sum_{c,y} P(c,y)\,\log \frac{P(c,y)}{P(c)P(y)}.$$

**Blockwise Flow.** The contribution of each pair $(a,b)$ is

$$C_{ab} = P(a,b)\,\log \frac{P(a,b)}{P(a)P(b)},$$

which we visualize as heatmaps, revealing which connections carry the most information.

## Results

We applied this methodology to a 3-layer MLP trained on MNIST (784–100–50–10), using manually chosen cluster counts to balance scalability to graphon representations with interpretability. The results are shown below.

### Graphon Visualizations

Figure 1 illustrates the weighted SBM graphons for each consecutive layer pair (L0→L1, L1→L2, L2→L3). The block structure highlights heterogeneity in connectivity. L0→L1 exhibits highly fragmented and uneven blocks, reflecting noisy interactions between raw pixels and the first hidden layer. L1→L2 shows clearer modular patterns, with strong positive and negative weight clusters, indicating the formation of higher–level features. L2→L3 reveals consolidated blocks, suggesting compression of earlier representations into task–aligned features.

### Information Bottleneck Dynamics

The resulting IB plane (Figure 2) demonstrates the expected trade-off between compression and prediction. Early layers (L0→L1) retain limited predictive information, consistent with diffuse clustering of raw inputs. Intermediate layers (L1→L2) achieve improved prediction while modestly increasing compression. The final layer (L2→L3) attains the highest balance of compression and predictive information, reflecting task-relevant feature extraction. This aligns well with the findings in (Shwartz-Ziv and Tishby, 2019)

### Block-Level Information Contributions

To further dissect the source of information, we computed the mutual information contributions of individual block–to–block connections (Figure 3). For L0→L1, contributions are sparse, with only a few block pairs carrying noticeable information, suggesting redundancy in raw features. For L1→L2, contributions are more distributed across multiple block pairs, indicating richer feature interactions. For L2→L3, contributions are highly concentrated, where a few dominant block pairs account for most of the predictive information, consistent with task-specific compression.

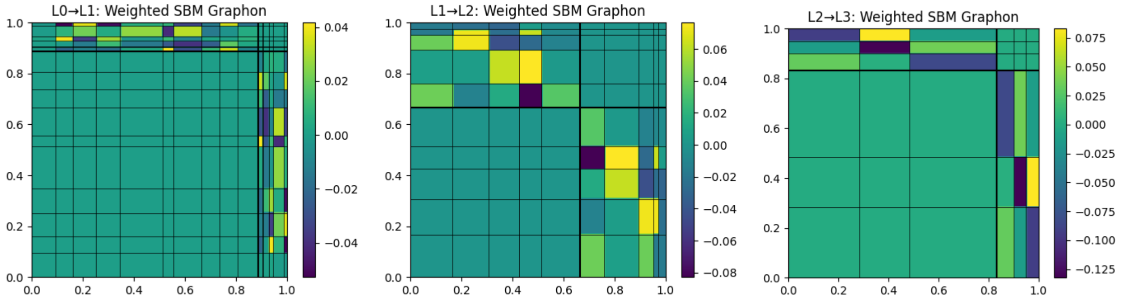

Figure 1: Weighted SBM graphons for each layer transition.

## 3. Conclusion and Outlook

Clustering neurons into block structures necessarily smooths weight patterns, and in the MNIST MLP setting this can oversimplify fine-grained dynamics. Nevertheless, as models scale, convergence toward graphon limits and averaging effects across larger clusters should yield more stable and informative estimates of mutual information. Preliminary experiments

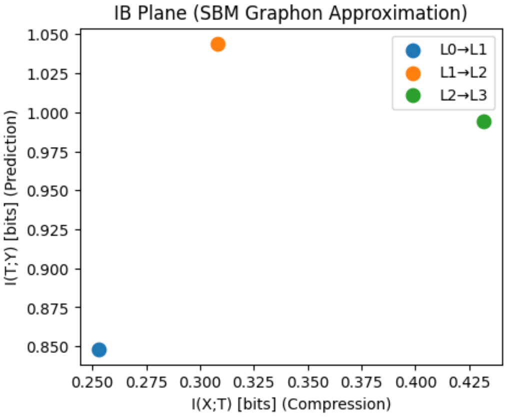

Figure 2: IB plane with compression $(I(X;T))$ vs. prediction $(I(T;Y))$.

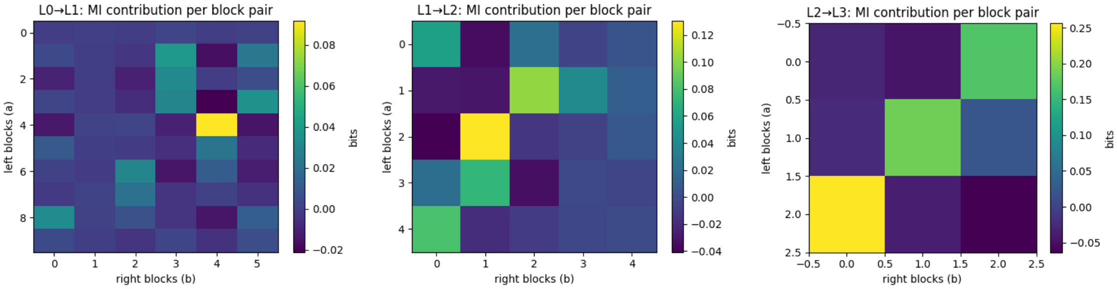

Figure 3: Mutual information contribution per block pair for each layer transition.

already suggest that blockwise MI patterns are not artifacts of random initialization but emerge through training, highlighting their potential explanatory value.

Future work will extend this framework to larger MLPs and transformer architectures, with the long-term goal of developing tractable graphon-based IB analyses for large language models (LLMs). Overall, our results suggest that graphon-based SBM approximations offer a promising path toward scalable, interpretable characterizations of information flow in deep networks.

## References

László Lovász. *Large Networks and Graph Limits*, volume 60 of *Colloquium Publications*. American Mathematical Society, 2012.

Thomas L. J. Ng and Thomas B. Murphy. Weighted stochastic block model. *Statistical Methods & Applications*, 30(5):1365–1398, 2021. doi: 10.1007/s10260-021-00578-4.

Ravid Shwartz-Ziv and Naftali Tishby. Opening the black box of deep neural networks via information. *Entropy*, 21(12):1181, 2019. doi: 10.3390/e21121181.

Chikashi Watanabe, Tatsuya Hara, and Hiroshi Nakagawa. Modular representation of layered neural networks. *Frontiers in Computational Neuroscience*, 11:71, 2017. doi: 10.3389/fncom.2017.00071.