NETWORK REGRESSION VIA SUPERVISED LATENT MOTIF

Yi Wei, Phoebe Kuang, Hanbaek Lyu†

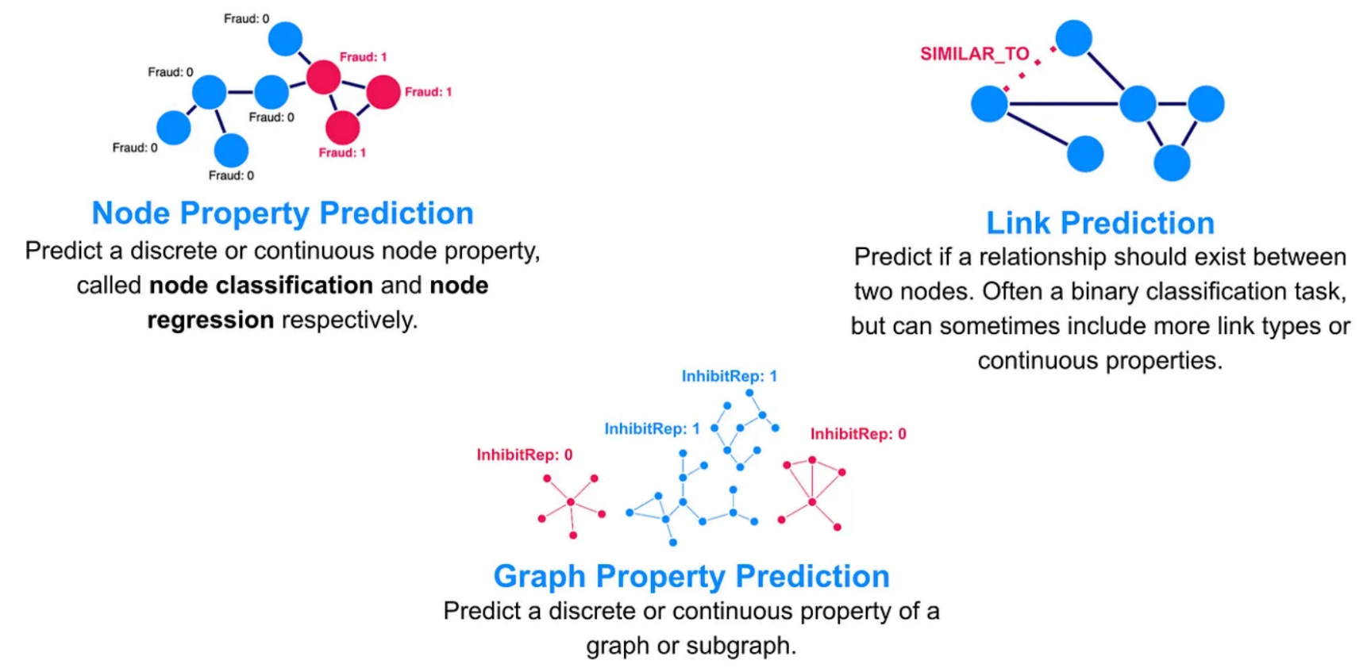
University of Wisconsin - Madison

Introduction

Networks are everywhere. Many complex systems—social, biological, technological—are naturally modeled as networks. Nodes represent entities (e.g., people, proteins, sensors) and edges represent interactions (e.g., friendships, chemical bonds, communication links).

Why network regression? Traditional regression predicts outcomes from vector inputs. In *network regression*, we predict outcomes or properties associated with the network structure itself, enabling us to:

- Identify which structural features (motifs, communities, hubs) drive certain network outcomes.
- Predict future links, node attributes, or global graph properties.



Node Property Prediction
Predict a discrete or continuous node property, called **node classification** and **node regression** respectively.

Link Prediction
Predict if a relationship should exist between two nodes. Often a binary classification task, but can sometimes include more link types or continuous properties.

Graph Property Prediction
Predict a discrete or continuous property of a graph or subgraph.

Tasks in Supervised Graph Machine Learning (Figure credit: Z. Blumenfeld)

Challenge:

- High-dimensional: Networks can be huge. Simple vectorization may lose structural information, and ad-hoc features can be too coarse.
- Complex Dependency. Unlike traditional regression models, which assume independence between observations, network data is defined by nodes connected through edges, introducing intricate dependencies that traditional models are not designed to handle.

Method: Supervised Network Dictionary Learning

Network Dictionary Learning Traditional Network Dictionary Learning (NDL) aims to represent complex networks as linear combinations of a small set of fundamental *motifs* or *subgraph patterns*, much like words form a dictionary. As shown in recent work [3], NDL successfully compresses large-scale networks into interpretable building blocks. Each dictionary element corresponds to a characteristic mesoscale pattern (e.g., a small induced subgraph) that recurs frequently throughout the network. By learning dictionaries from real-world networks (e.g., university Facebook networks), we can visualize systematic differences in structural patterns across them.

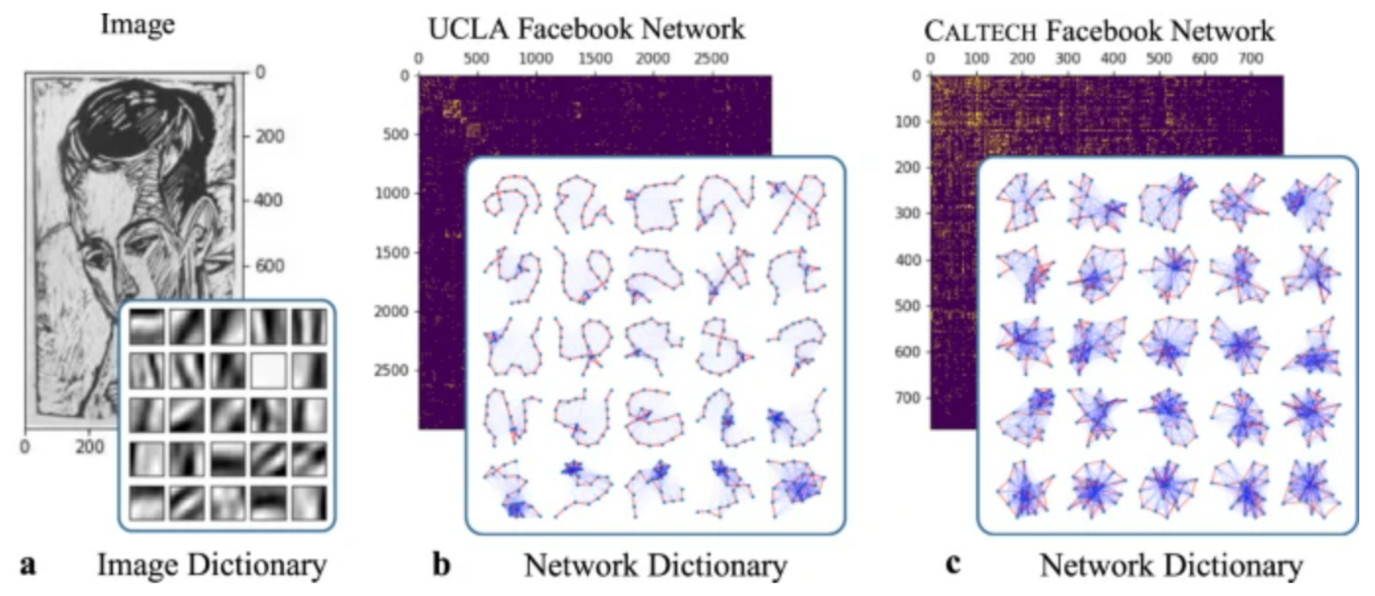


Illustration of mesoscale structures that we learn from images and networks.[3]

However, NDL alone focuses primarily on reconstructing the network structure and does not incorporate external labels or responses. In many applications, we need to not only understand the structure but also predict outcomes associated with the network—such as node attributes, link formation likelihood, or global network-level responses.

Enter SNDL: Adding Supervision. Supervised Network Dictionary Learning (SNDL) augments NDL with a predictive component. We introduce a response variable Y (e.g., class labels, continuous attributes) associated with one or more networks. SNDL learns a dictionary of motifs *and* simultaneously ensures that these motifs are predictive of Y. Formally, given a collection of d networks $\{G_i\}$ with labels $\{y_i\}$, SNDL seeks a factorization:

$$\min_{W,H,\beta} \xi\|\mathbf{X} - WH\|_F^2 + \sum_{i=1}^{n} \ell(y_i, \beta^\top W^\top x_i),$$

where $\mathbf{X}$ is a feature matrix constructed from sampled k-node subgraphs (motifs) of the input networks, W is the dictionary matrix whose columns represent latent motifs, H encodes how each motif is used to reconstruct $\mathbf{X}$, and β links the motifs to the response Y, ensuring that learned motifs are informative predictors.

Sampling Mesoscale Structure. To populate $\mathbf{X}$, we must sample representative k-node subgraphs. We employ a k-path sampling algorithm (e.g., pivot chain sampling), which efficiently generates connected k-node subgraphs uniformly at random [2]. After effective sampling, we vectorized adjacency matrices of sampled subgraphs into feature vectors, and optionally concatenate node/edge features (if any). Stacking these feature vectors for all samples from all networks yields our feature matrix $\mathbf{X} \in \mathbb{R}^{p \times n}$. The associated label vector $\mathbf{Y}$ is formed by repeating y_i for all samples taken from G_i.

Optimization via Block Coordinate Descent (BCD). The SNDL objective is nonconvex, involving a supervised loss term (logistic regression on the motifs) plus a matrix factorization term. Recent theoretical results [1] show that a suitably designed BCD algorithm converges to an ϵ-stationary point in polynomial time. The algorithm alternates updates over W, H, and β, each step projecting onto convex constraint sets (e.g., nonnegativity) and using adaptive step sizes.

Theory: Global Reconstruction Accuracy

Global Accuracy Bounds. A key theoretical insight is that controlling the SNDL objective simultaneously ensures both accurate network reconstruction and good predictive performance. More concretely, suppose we have an optimal solution (W, H, β) minimizing:

$$F(W, H, \beta) = \xi\|\mathbf{X} - WH\|_F^2 + \sum_{i=1}^{n} \ell(y_i, \beta^\top W^\top x_i).$$

Then, the total error—measured as the expectation of the network reconstruction error and the prediction error—is bounded by the minimum value of F. Intuitively, if we find a factorization and regression coefficients that both approximate the observed subgraphs well and classify their labels accurately, we can guarantee global-level performance with respect to local error.

Network Reconstruction Error Bound. If our learned dictionary motifs approximate the sampled k-node subgraphs well, then the reconstructed adjacency matrix $\hat{A}$ matches A with small error at the global scale. Formally we obtain:

$$d_{JD}(G, G_{\text{recons}}) \leq \frac{1}{k}\mathbb{E}_{\mathbf{x}}[\|A_{\mathbf{x}} - \hat{A}_{\mathbf{x};W}\|_1],$$

where d_{JD} measures a Jaccard-type distance between the original network G and its reconstruction G_{recons}.

Prediction Error Bound. Similarly, incorporating supervised loss terms ensures that the dictionary not only explains network structure but is also informative for downstream prediction tasks. Theoretically, we can show that the prediction error, as measured by a multinomial logistic loss, shrinks as the SMF objective is minimized. Hence, good solutions yield small classification (or regression) errors.

Combined Guarantee. Together, these results imply:

$$d_{JD}(G, G_{\text{recons}}) + D_{KL}(Q\|P) = C + \mathbb{E}_{\mathbf{x}}[\ell(\mathbf{y}, \mathbf{a})] + \frac{1}{k}\mathbb{E}_{\mathbf{x}}[\|A_{\mathbf{x}} - \hat{A}_{\mathbf{x};W}\|_1]$$

Where C is some constant, P is the label probability predictive distribution with respect to the learned dictionary and the learned logistic regression coefficient, Q is the label distribution, $\mathbf{y}$ is the real label vector, and $\mathbf{a}$ is the filter-based activation.

Applications

Synthetic Data: Validating the Approach. We tested our model on synthetic networks generated from random graph models, including Erdős-Rényi (ER), Barabási-Albert (BA), Watts-Strogatz (WS), and the Configuration Model (CM). By applying SNDL to discriminate between a real network and a synthetic one, we demonstrated that the learned dictionary motifs capture crucial structural differences (Fig 1.).

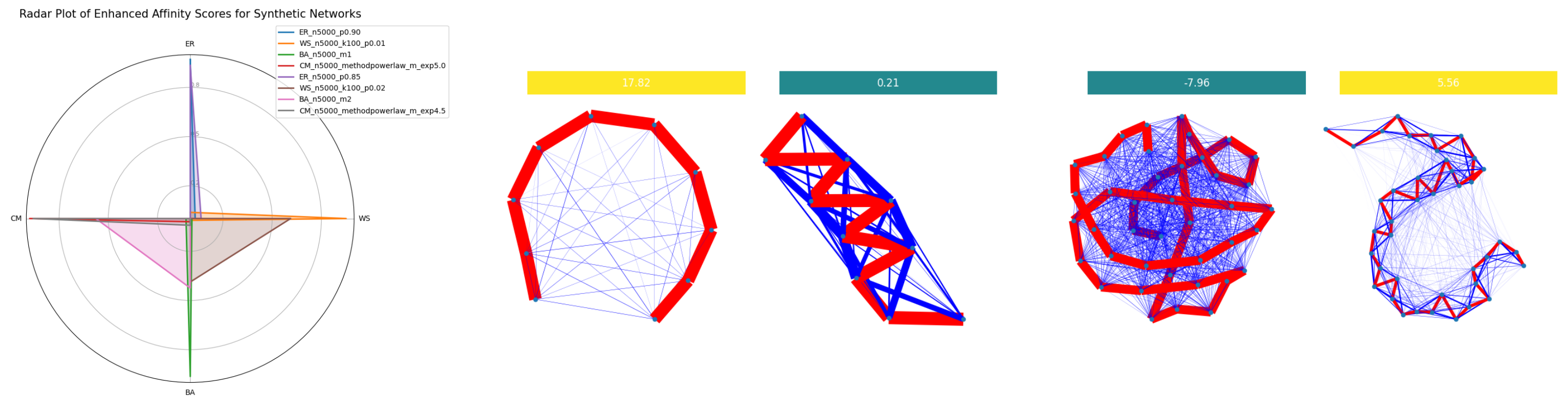


Fig 1. Synthetic Network Validation

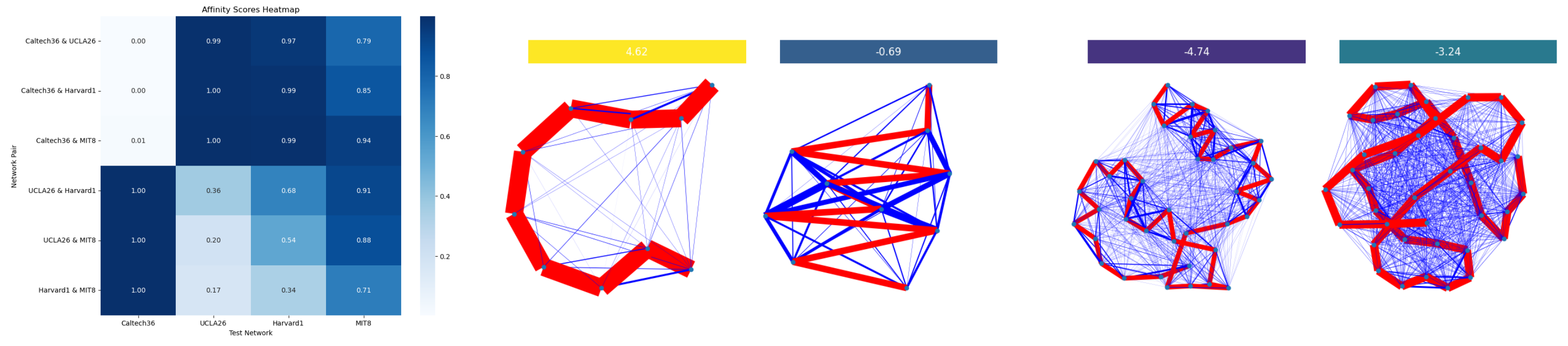
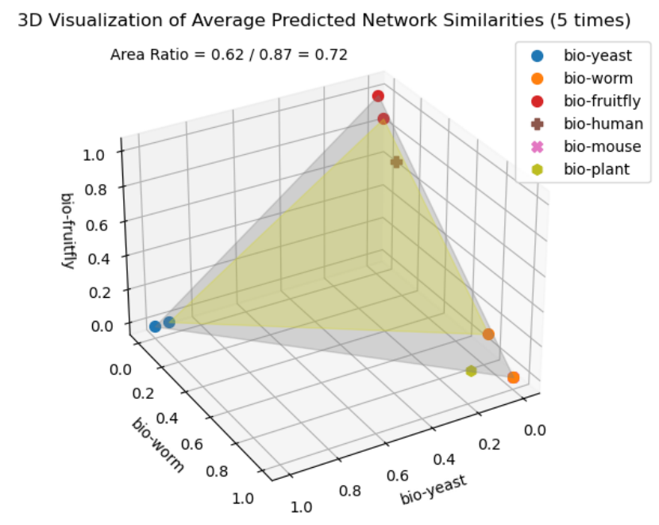


Fig 2. Social Network Example

Real-World Networks: Social and Biological Insights. Moreover, we applied SNDL to real-world networks, focusing on social and biological domains. Using the Facebook100 dataset [4], we employed SNDL to predict which university a newly observed subgraph likely came from. The learned dictionaries revealed interpretable mesoscale motifs that differentiate some institutions (e.g., Caltech) from others with more complex structures (e.g., Harvard, UCLA and MIT) (Fig 2.).



Network	Coordinates	Std Dev
bio-yeast	(0.95, 0.02, 0.02)	(0.00, 0.00, 0.00)
bio-worm	(0.01, 0.84, 0.15)	(0.00, 0.01, 0.01)
bio-fruitfly	(0.02, 0.07, 0.91)	(0.00, 0.01, 0.01)
bio-human	(0.04, 0.21, 0.75)	(0.01, 0.02, 0.02)
bio-mouse	(0.00, 1.00, 0.00)	(0.00, 0.00, 0.00)
bio-plant	(0.11, 0.88, 0.00)	(0.13, 0.12, 0.00)

We also experimented with Protein-Protein Interaction (PPI) networks from BioGRID[4]. Here, the dictionary motifs learned by SNDL shed light on how biological networks differ across species (e.g., yeast vs. human vs. fruitfly, etc.). This approach can guide researchers toward identifying structural motifs linked to specific biological functions or conditions.

Across these diverse datasets—synthetic SNDL consistently extracted interpretable dictionaries and provided meaningful predictions, demonstrating its versatility and potential for broad application in network science.

References

[1] Joowon Lee, Hanbaek Lyu, and Weixin Yao. “Supervised Matrix Factorization: Local Landscape Analysis and Applications”. In: *Proceedings of the 41st International Conference on Machine Learning*. Vol. 235. Proceedings of Machine Learning Research. PMLR, 21–27 Jul 2024, pp. 26752–26788. URL: `https://proceedings.mlr.press/v235/lee24p.html`.

[2] Hanbaek Lyu, Facundo Memoli, and David Sivakoff. “Sampling random graph homomorphisms and applications to network data analysis”. In: *Journal of Machine Learning Research* 24.9 (2023), pp. 1–79. URL: `http://jmlr.org/papers/v24/20-449.html`.

[3] Hanbaek Lyu et al. “Learning low-rank latent mesoscale structures in networks”. In: *Nature Communications* 15.1 (Jan. 2024). DOI: `10.1038/s41467-023-42859-2`.

[4] Ryan A. Rossi and Nesreen K. Ahmed. “The Network Data Repository with Interactive Graph Analytics and Visualization”. In: *AAAI*. 2015. URL: `https://networkrepository.com`.