

Interpretable and Parameter Efficient Graph Neural Additive Models with Random Fourier Features

Thummaluru Siddartha Reddy

Vempalli Nagasai Saketh

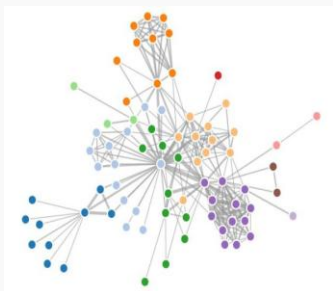
Mahesh Chandran

Fujitsu Research of India, Bangalore

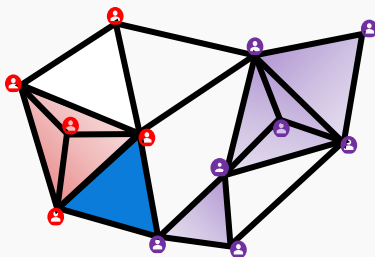


Graphs and Graph Neural Networks

- ❖ **Graphs** are mathematical objects that captures the relationship between the entities (nodes) through edges.



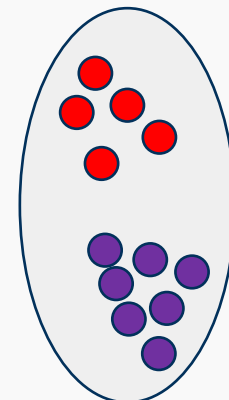
Social networks



Citation networks



Graph Neural Networks

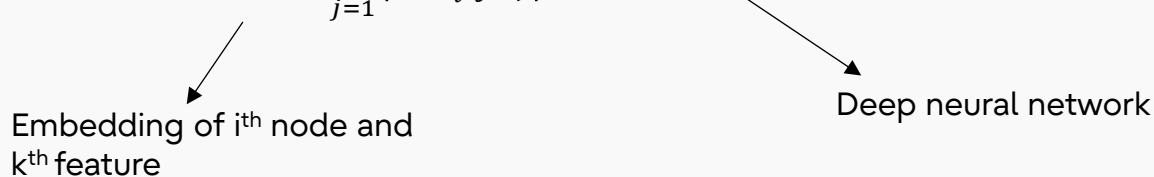


Embedding space

- ❖ GNNs are **deep representation learning** architectures with **graphs** as inductive bias
- ❖ GNNs obtains the embeddings using the message passing or graph convolution techniques which **typically mixes features**. Therefore the representations obtained from GNNs lack interpretability.

- ❖ Additive modelling is workhorse principle behind interpretable deep learning architectures
- ❖ GNAN is the first graph neural additive model that models contributions of each features using a DNN.

Set up : Graph (G) with N nodes and feature matrix as $\mathbf{X} \in R^{N \times D}$

$$z_{i,k} = \sum_{j=1}^N \frac{1}{|dist_i(j,i)|} \rho(dist(i,j)) f_k(x_{i,j})$$


Embedding of i^{th} node and k^{th} feature

Deep neural network

Challenges

- ❖ Training a DNN for each feature is computationally intense if feature dimension is large.
- ❖ Computing a `dist()` function costs $O(N^2 + NE)$, which and is not feasible for large scale graphs.


Motivation : To propose a **light-weight graph model** that is **inherently interpretable**

Set up : Graph (G) with N nodes and feature matrix as $\mathbf{X} \in R^{N \times D}$

Framework : We model each feature contribution with a **GP prior** whose kernel admits **RFF approximation**.

$$y_i = \sum_{k=1}^D f_k(x_{i,k})$$

$$f_k = \mathcal{GP}(0, \mathbf{K}_G(.))$$


$$\mathbf{K}_G(.) = \frac{a_{i,j}(x_{i,k} - x_{j,k})^2}{2\theta^2}$$

- ❖ The kernel can be equivalently expressed as $\mathbf{K}_G(.) = \frac{(\tilde{x}_{i,k} - \tilde{x}_{j,k})^2}{2\theta^2}$
- ❖ Proposed kernel is **shift invariant and positive definite** for a fixed node pair kernel can be approximated as $\mathbf{K}_G(.) = \Phi_a^T(x_{i,k})\Phi_a(x_{j,k})$

Drawback: It captures only information from **one-hop** and fails to aggregate information from **multi-hops**

Graph Neural Additive Model with RFFs

- ❖ To obtain the information from **multi-hop** we process the data using **FIR filter**

$$\mathbf{H} = \sum_{h=0}^R \alpha_h \mathbf{L}_G^h$$

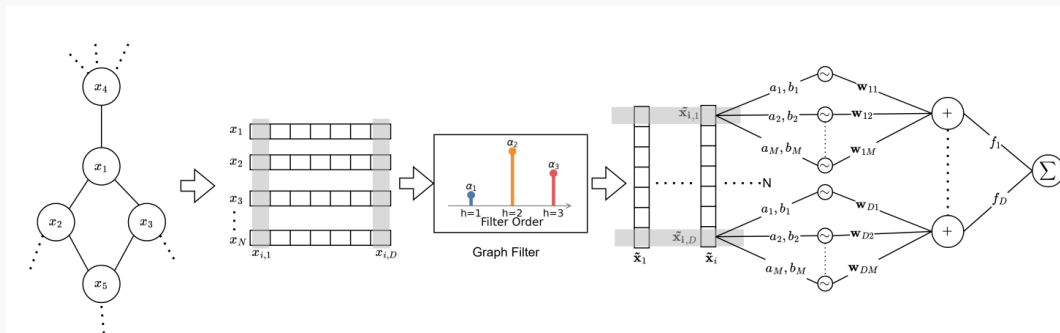
Learnable filter coefficients

Normalized graph Laplacian

- ❖ The **filtered features** are given as $\tilde{\mathbf{X}} = \mathbf{H}\mathbf{X}$

- ❖ Now the kernel is modified to account for the information from multi-hops $\mathbf{K}_G(\cdot) = \frac{\{\tilde{x}_{i,k} - \tilde{x}_{j,k}\}^2}{2\Theta^2}$ $\left| \mathbf{K}_G(\cdot) = \Phi_a^T(x_{i,k})\Phi_a(x_{j,k}) \right.$

- ❖ Leveraging the kernel approximation GP prior now transforms as **Bayesian prior** as $f_k(x_{i,k}) = \Phi_a^T(\tilde{x}_{i,k})\mathbf{w}_k$

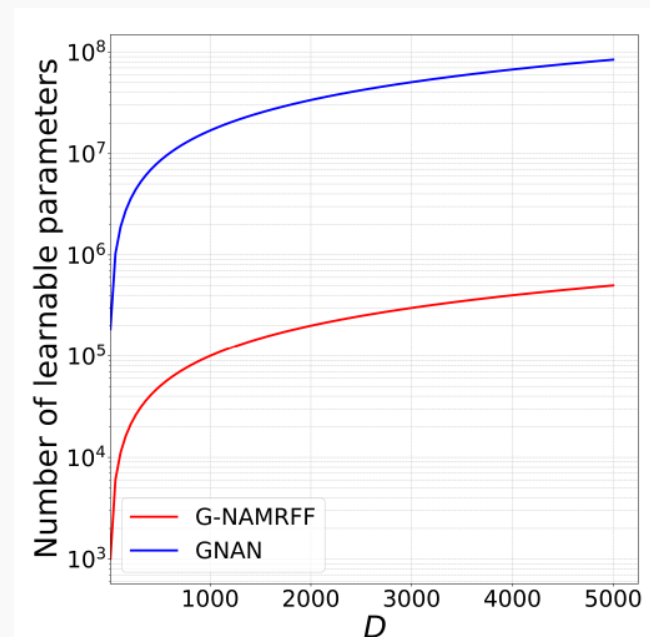


$$y_i = \sum_{k=1}^D \Phi_a^T(\tilde{x}_{i,k})\mathbf{w}_k$$

❖ Number of learnable parameters

- GNAN: $(D + 1) \times (H_u^2 \times (L - 1) + (L + 2) \times H_u + 1)$
 - G-NAMRFF: $D \times M + R + 1$
- 168x fewer parameters compared to GNAN

$$H_u = 64, L = 5, D = 100, M = 100, R = 5$$



Permutation equivariance:

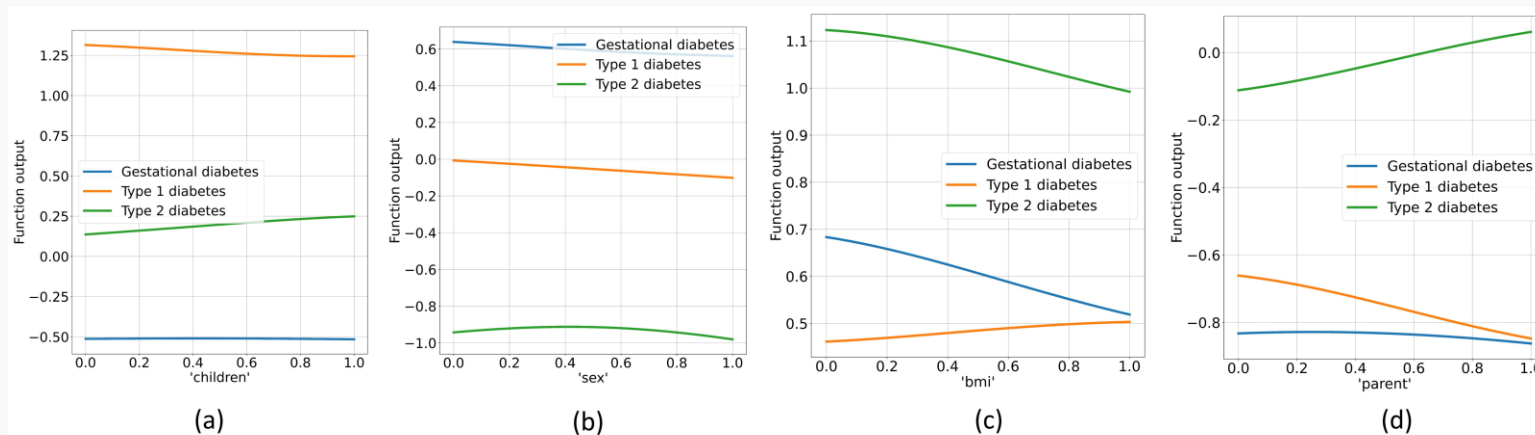
Theorem (Permutation Equivariance). Let $\mathcal{P} = \{ \mathbf{P} \in \{0, 1\}^{N \times N} : \mathbf{P}^\top \mathbf{P} = \mathbf{P} \mathbf{P}^\top = \mathbf{I}_N \}$ be the set of all $N \times N$ permutation matrices. Then under the permutation of the graph Laplacian $\mathbf{L}_{\mathcal{G}}$ and node-feature matrix \mathbf{X} by any $\mathbf{P} \in \mathcal{P}$, the predictions from \mathcal{G} -NAMRFF also modifies as $\mathbf{y}_{perm} = \mathbf{P} \mathbf{y}$, where $\mathbf{y} \in \mathbb{R}^N$ is the predictions across all the nodes.

Robustness to perturbations:

Theorem (Robustness to perturbation of graph Laplacian). Let $\mathbf{L}_{\mathcal{G}} = \mathbf{L}_{\mathcal{G}} + \Delta \mathbf{L}_{\mathcal{G}}$ be the Laplacian of the perturbed graph, with $\|\Delta \mathbf{L}_{\mathcal{G}}\|_2 \leq \epsilon$, and assume that the RFF map $\Phi_a(\cdot)$ is C_{RFF} -Lipschitz continuous. Then each node prediction satisfies $|\hat{y}_i - y_i| \leq C K \epsilon D \|\mathbf{X}\|_2$, where $C = C_{\text{RFF}} (\max_k \|w_k\|_2)$, $K = \frac{1}{4} \|\boldsymbol{\alpha}\|_1 (R^2 - 1) \left(\frac{R-1}{R+1} \right)^R$ are constants.

Interpretability on Node Classification Task

Pubmed dataset: Multiclass citation-network dataset where nodes are research articles categorized into three diabetes classes



Mutagenicity Dataset



Empirical Results

Model	Cora	Citeseer	Pubmed	Cornell	ogbn-arxiv	ogbn-products
GCN [19]	81.23 \pm 1.1	71.20 \pm 1.7	78.50 \pm 1.3	65.90 \pm 0.5	71.74 \pm 0.3	75.64 \pm 0.3
GAT [30]	80.32 \pm 2.3	70.26 \pm 2.3	77.12 \pm 2.4	72.50 \pm 0.7	71.95 \pm 0.6	79.45 \pm 0.5
GraphSAGE [12]	79.94 \pm 3.4	65.12 \pm 1.9	78.25 \pm 1.2	75.90 \pm 5.0	71.49 \pm 0.2	75.63 \pm 0.3
Graph Transformer [37]	80.70 \pm 0.5	76.00 \pm 0.9	78.80 \pm 1.4	70.50 \pm 1.7	70.13 \pm 0.5	74.74 \pm 0.5
NAM [2]	51.35 \pm 2.3	55.40 \pm 1.9	58.16 \pm 2.3	59.15 \pm 2.6	56.12 \pm 3.4	OOM
GPAM [39]	59.96 \pm 3.2	60.30 \pm 3.9	62.30 \pm 3.7	60.12 \pm 3.6	62.35 \pm 4.2	60.13 \pm 3.9
GNAN [5]	77.89 \pm 5.1	65.23 \pm 3.7	75.13 \pm 2.4	71.76 \pm 4.2	69.56 \pm 0.9	OOM
G-NAMRFF ($R = 1$)	75.32 \pm 1.8	67.12 \pm 1.1	75.12 \pm 3.8	64.12 \pm 2.9	66.94 \pm 1.6	55.73 \pm 1.8
G-NAMRFF	79.84 \pm 1.7	69.45 \pm 2.5	77.30 \pm 1.4	73.54 \pm 4.9	70.02 \pm 3.9	72.13 \pm 0.4

Node classification task

Model	Proteins	Mutag	Mutagenicity	NCI1	PTC
GCN [19]	70.97 \pm 4.6	68.07 \pm 6.3	75.69 \pm 0.9	66.35 \pm 1.3	56.98 \pm 5.8
GAT [30]	69.92 \pm 4.0	67.20 \pm 3.4	69.40 \pm 1.2	66.12 \pm 2.1	55.60 \pm 11.1
GraphSAGE [12]	67.35 \pm 2.3	64.12 \pm 2.4	69.25 \pm 3.9	65.56 \pm 3.9	57.12 \pm 4.9
Graph Transformer [37]	69.76 \pm 3.2	66.30 \pm 5.3	73.10 \pm 0.9	68.24 \pm 3.4	55.90 \pm 3.5
NAM [2]	62.45 \pm 4.2	63.12 \pm 9.1	67.35 \pm 2.5	57.15 \pm 1.2	54.97 \pm 7.5
GPAM [39]	65.68 \pm 4.1	64.30 \pm 8.4	65.46 \pm 2.2	53.80 \pm 2.7	52.65 \pm 8.1
GNAN [5]	59.64 \pm 2.4	67.35 \pm 3.9	66.64 \pm 4.7	50.87 \pm 1.4	55.07 \pm 5.2
G-NAMRFF ($R=1$)	67.83 \pm 4.4	71.20 \pm 6.7	68.98 \pm 3.4	63.65 \pm 2.8	55.65 \pm 5.6
G-NAMRFF	69.94 \pm 3.7	79.81 \pm 5.3	71.70 \pm 2.0	66.10 \pm 1.7	61.91 \pm 3.4

Graph classification task

