LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Universality, Approximation Rates, and Generalization Bounds of PENNs

Thilo Meyer-Brandis
Niklas Weber

Joint work with Lukas Gonon

Swissquote Workshop on Graph Learning in Financial Networks (EPFL)

Permutation Equivariant Functions

- Domain of directed graphs with $N \in \mathbb{N}$ nodes and $d, d' \in \mathbb{N}$ dimensional node and edge features $\mathcal{D}_{N,d,d'} := \mathbb{R}^{N \times d} \times \mathbb{R}^{N \times N \times d'}$
- A function $\tau : \mathcal{D}_{N,d,d'} \to \mathbb{R}^{N \times l}$ is permutation equivariant if

$$\tau(\sigma(a, \ell)) = \sigma(\tau(a, \ell)),$$

  for any $(a, \ell) \in \mathcal{D}_{N,d,d'}$ and permutation σ of $[N]$.
- The permutation $\sigma(a, \ell)$ of a graph $(a, l) \in \mathcal{D}_{N,d,d'}$ is defined as $(\sigma(a), \sigma(\ell))$ where

$$\sigma(a)_{i,k} = a_{\sigma^{-1}(i),k} \qquad \text{for } i \in [N], k \in [d]$$
$$\sigma(\ell)_{i,j,k} = \ell_{\sigma^{-1}(i),\sigma^{-1}(j),k} \qquad \text{for } i, j \in [N], k \in [d']$$

Permutation Equivariant Neural Networks (PENNs)

- We propose approximating permutation equivariant functions by **permutation equivariant neural networks (PENNs)**.
- A PENN : $\mathcal{D}_{N,d,d'} \to \mathbb{R}^{N \times l}$ is defined by

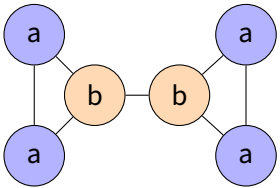
$$\mathrm{PENN}(a, \ell)_k := \hat{\rho}\left(a_k, \sum_{j \in \mathcal{N}(k)} \hat{\psi}(a_k, \ell_{jk}, a_j), \sum_{i=1}^{N} \hat{\alpha}\left(a_i, \sum_{j \in \mathcal{N}(i)} \hat{\varphi}\left(a_i, \ell_{ji}, a_j\right)\right)\right).$$

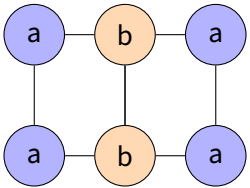
  for FNNs $\hat{\rho}$, $\hat{\alpha}$, $\hat{\psi}$, and $\hat{\varphi}$.
- PENNs belong to the family of **message passing GNNs** and cover many commonly used GNN architectures.

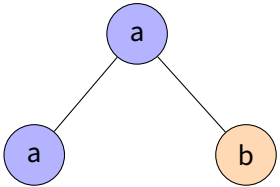
Outline

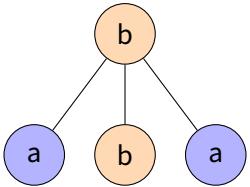
Properties of Message Passing

- Message passing is **permutation equivariant**
- Simple Message passing GNNs are **not universal** (tree-like equivalence)









Approaches to improve the expressiveness of GNNs

- Considering **k-hop neighborhoods** (Morris et al. [9], Maron et al. [7]) or **high-dimensional tensors representations** (Maron et al. [6], Maron et al. [8], Keriven and Peyré [5]) $\rightarrow$ computational inefficient.
- Alternative: enrich node features and still use standard GNNs
- Adding **unique node labels** (Murphy et al. [10], Dasoulas et al. [3]).
- Adding **random node features** (Sato et al. [12]).
- With random node features there exist few universality results with limitations:
  - Invariant functions on simple graphs (undirected, no node or edge features) of fixed size (Abboud et al. [1]) or continuous invariant functions on bounded domains of directed graphs with node and 1-dimensional edge features (Puny et al. [11]).
- We improve on all these results and show universal approximation for measurable in- and equivariant functions on directed graphs with multidimensional node and edge features.

Random Node Features

1. We extend the feature space by additional d_r-dimensional node features with values in $R^{N \times d_r}$.
2. Universality in probability $\Longrightarrow$ we consider random graphs $(A, L) : \Omega \to \mathcal{D}_{N,d,d'}$ and random additional node features $R : \Omega \to R^{N \times d_r}$ on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

Definition

We say that random features $R : \Omega \to \mathbb{R}^{N \times d_r}$ provide finite unique node features with high probability if for any $p \in (0, 1)$ there exist $M \in \mathbb{N}$ and a function $h : \mathbb{R}^{d_r} \to [M]$ such that

$$\mathbb{P}\left(\forall i, j \in [N], i \neq j : h(R_i) \neq h(R_j)\right) \geq p.$$

Examples for finite unique node features w.h.p.

- Deterministic features, e.g. $R = (1, \dots, N)$
- One-hot or binary encoding of node IDs

$$R = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ & & \ddots & \\ 0 & 0 & \dots & 1 \end{pmatrix}, R = \begin{pmatrix} 0 & 0 & \dots & 0 \\ 0 & 0 & \dots & 1 \\ & & \iddots & \\ 1 & 1 & \dots & 1 \end{pmatrix}$$

- Random permutations $\sigma(R)$ of any deterministic R
- $R = (R_1, \dots, R_N) \in L^0(\mathbb{R}^N)$ with continuous i.i.d. components R_i, e.g. uniformly on $[0, 1]$ or normally on $\mathbb{R}$.

Universality of R-PENNs

- For a PENN : $\mathcal{D}_{N,d+d_r,d'} \to \mathbb{R}^{N\times l}$ on the extended domain $\mathbb{R}^{N\times d+d_r} \times \mathbb{R}^{N\times N\times d'}$ and random node features $R : \Omega \to R^{N\times d_r}$, we define

$$\text{R-PENN}(a,\ell) := \text{PENN}((a,r),\ell)|_{r=R} \qquad (a,\ell) \in \mathcal{D}_{N,d,d'}$$

Theorem

Let $(A, L) : \Omega \to \mathcal{D}_{N,d,d'}$ be any random graph and $R : \Omega \to \mathbb{R}^{N\times d_r}$ any finite unique node features with high probability.
Further, let $\tau : \mathcal{D}_{N,d,d'} \to \mathbb{R}^{N\times l}$ be a measurable, permutation equivariant function.
Then, for $\varepsilon, \delta > 0$, there exists an R-PENN *such that*

$$\mathbb{P}\left(\|\tau((A,L)) - \text{R-PENN}(A,L)\| \leq \varepsilon\right) \geq 1-\delta.$$

Are R-PENNs still permutation equivariant?

- Adding random node features destroys permutation equivariance of the R-PENN on $\mathcal{D}_{N,d,d'}$.
- In general, for a permutation σ

$$\sigma(\text{R-PENN}(a, \ell)) \neq \text{R-PENN}(\sigma(a, \ell))$$

- However, we can recover the following:

Proposition

If $\sigma(R)$ and R are distributed identically for a permutation $\sigma \in S_N$, then $\sigma(\text{R-PENN}(a, \ell))$ and $\text{R-PENN}(\sigma(a, \ell))$ are distributed identically for any $(a, \ell) \in \mathcal{D}_{N,d,d'}$. In particular, assuming integrability, we obtain permutation equivariance in expectation:

$$\mathbb{E}\left[\sigma(\text{R-PENN}(a, \ell))\right] = \mathbb{E}\left[\text{R-PENN}(\sigma(a, \ell))\right].$$

- Assume $\sigma(R) \sim R$ for all permutations $\sigma \in S_N$.
- For i.i.d. copies $R^{(1)}, \ldots, R^{(M)}$ of R, by the law of large enumbers we get for any $\sigma \in S_N$ and $(a, \ell) \in \mathcal{D}_{N,d,d'}$

$$\lim_{M\to\infty} \frac{1}{M} \sum_{i=1}^{M} \sigma(\mathrm{R^i\text{-}PENN}(a, \ell)) = \lim_{M\to\infty} \frac{1}{M} \sum_{i=1}^{M} \mathrm{R^i\text{-}PENN}(\sigma(a, \ell)).$$

- Therefore, with the class of functions

$$\frac{1}{M} \sum_{i=1}^{M} \mathrm{R^i\text{-}PENN}((a, \ell) \qquad\qquad (a, \ell) \in \mathcal{D}_{N,d,d'}$$

we can “increase” permutation equivariance by considering an average over $M \in \mathbb{N}$ independent copies of R. This can improve performance, e.g., faster convergence and improved generalization as in Bechler-Speicher et al. [2].

Outline

Stronger Assumptions

We make the following (stronger) assumptions in order to establish approximation rates:

- We consider the **compact** domain $\mathcal{D}_{N,d,d'}^{[0,1]} := [0,1]^{N\times d} \times [0,1]^{N\times N\times d'}$.
- The random features $R : \Omega \to [0,1]^N$ provide finite unique node features and have support on $[0,1]^N$.
- $\tau : \mathcal{D}_{N,d,d'} \to \mathbb{R}^{N\times l}$ is a $k \in \mathbb{N}_{\geq 2}$ times **continuously differentiable** permutation equivariant function.

Approximation Rates of R-PENNs

Theorem

Assume above assumptions. Then, for every $\delta \in (0,1)$ there exist constants c_1, $c_2 > 0$ such that for every $\varepsilon > 0$ there exists an R-PENN with

$$\mathbb{P}\left(\max_{(a,\ell)\in\mathcal{D}_{N,d,d'}^{[0,1]}} \|\tau(a,\ell) - \text{R-PENN}(a,\ell)\| \leq \varepsilon \right) \geq 1-\delta.$$

Moreover, we have the following upper bounds of the neural network complexity:
For $\chi \in \{\hat{\rho}, \hat{\alpha}, \hat{\phi}, \hat{\psi}\}$ we have $L(\chi) \leq c_1 \log_2\left(\frac{1}{\varepsilon}\right)$, and $M(\chi) \leq c_1 \left(\frac{1}{\varepsilon}\right)^{c_2/(k-1)} \left(\log_2\left(\frac{1}{\varepsilon}\right)\right)^2$.

- $\hat{\rho}, \hat{\alpha}, \hat{\varphi}, \hat{\psi}$ are the FNNs specifying the R-PENN.
- $L(\chi)$ is the number of layers and $M(\chi)$ the number of non zero weights.
- c_2 is explicitly goven in terms of N, d, d', and δ

Outline

Generalization Gap

- Let $\mathcal{G}$ be a set of graphs and $\mathcal{Y}$ a set of labels.
- Let $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{G}}$ be a hypothesis class (later: class of PENNs).
- Let $\mathcal{Z} := \mathcal{G} \times \mathcal{Y}$ be the space of graph–label pairs, with distribution μ.
- Let $\mathcal{S} = (G_1, Y_1), \ldots, (G_k, Y_k)) \sim \mu^k$ be an (i.i.d.) *random sample* of size k
- A realization $s = ((g_1, y_1), \ldots, (g_k, y_k)) \in \mathcal{Z}^k$ is sample of size k.
- A graph learning algorithm $\mathcal{A} : \mathcal{Z}^k \to \mathcal{H}$ assigns to each sample s a hypothesis $\mathcal{A}_s$.
- Let $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ be a bounded loss function.
- The expected and empirical errors of $\mathcal{A}_s$ are defined as

$$\ell_{\exp}(\mathcal{A}_s) := \mathbb{E}_{(G,Y)\sim\mu}[\ell(\mathcal{A}_s(G), Y)], \qquad \ell_{\mathrm{emp}}(\mathcal{A}_s) := \frac{1}{|s|} \sum_{(g,y)\in s} \ell(\mathcal{A}_s(g), y).$$

- The quantity $|\ell_{\exp}(\mathcal{A}_s) - \ell_{\mathrm{emp}}(\mathcal{A}_s)|$ is the *generalization gap*.

Road to Generalization Bounds

- Introduce pseudo-metric on graph space (via unrolling tree).
- Show Lipschitz continuity of all function in $\mathcal{H}$ with respect to pseudo metric.
- Robustness setting of Xu and Mannor [13] yields generalization bounds depending on covering number.
- Bounding covering number yields actual bounds.

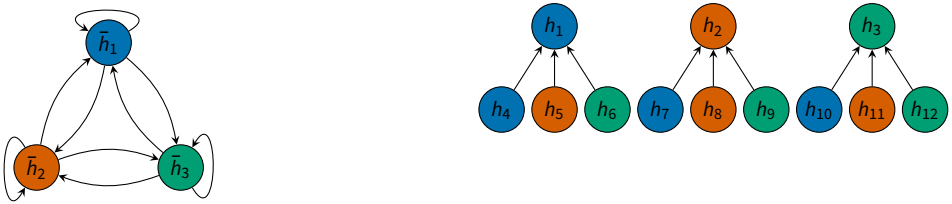


Figure: Graph X (left) and Level 1 unrolling tree $\mathcal{T}_1(X)$ (right). Note: $\bar{h}_1 = h_1 = h_4 = h_7 = h_{10}$, $\bar{h}_2 = h_2 = h_5 = h_8 = h_{11}$, and $\bar{h}_3 = h_3 = h_6 = h_9 = h_{12}$.

Pseudo-metrics

Definition

Let $L \in \mathbb{N}_0$, $X, Y \in \mathcal{G}_N$, and $\mathcal{T}_L(X)$, $\mathcal{T}_L(Y)$ the level L unrolling-trees of X, Y, respectively. For $i, j \in V(\mathcal{T}_L(X))$, let $h_i^X, e_{i,j}^X$ and $h_i^Y, e_{i,j}^Y$ denote the node and edge features of $\mathcal{T}_L(X)$ and $\mathcal{T}_L(Y)$, respectively. We define

$$d_L(X,Y) := \min_{\substack{\chi: V(\mathcal{T}_L(X)) \to V(\mathcal{T}_L(Y)) \\ \text{edge-preserving bijection}}} \sum_{\substack{i \in V(\mathcal{T}_L(X)) \\ j = \mathrm{par}(i)}} \left\| (h_j^X,\, e_{i,j}^X,\, h_i^X) - (h_{\chi(j)}^Y,\, e_{\chi(i),\chi(j)}^Y,\, h_{\chi(i)}^Y) \right\|, \tag{1}$$

where we set $\chi(0) = 0$, $h_0^X = h_0^Y = \mathbf{0}$, and $e_{\cdot,0}^X = e_{\cdot,0}^Y = \mathbf{0}$.

This pseudo-metric is the smallest total euclidean distance between (node, edge, node)-triples of two unrolling trees that can be obtained among all edge-preserving bijections.

Edge-Preserving Bijection

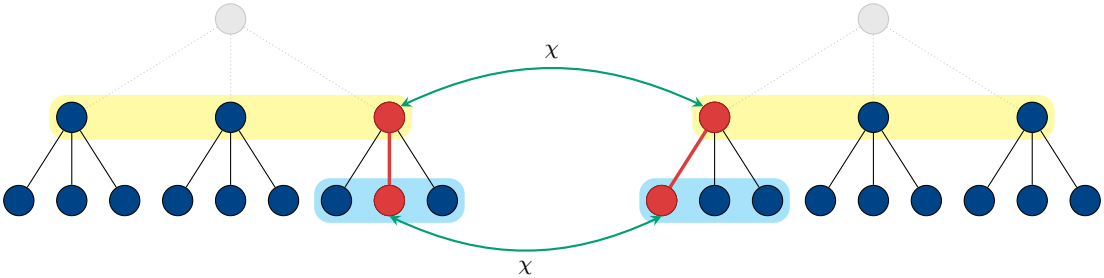


Figure: Two trees with virtual roots (grey). The bijection χ maps main nodes of the left tree to main nodes of the right tree (yellow regions), and correspondingly their children (cyan regions); shown for one highlighted pair in red.

PENN Class

- Let $X \in \mathcal{G}_N$ and $L \in \mathbb{N}$.
- Let ρ, α, ϕ, and ψ be functions of appropriate dimensions. We could also have different ρ, α, ϕ, ψ for different layers.
- In each message-passing step l, for each k:

$$h_k^{(l)} = \rho\left(h_k^{(l-1)}, \sum_{i \in \mathcal{N}(k)} \phi\left(h_k^{(l-1)}, e_{i,k}, h_i^{(l-1)}\right), \sum_{j \in [N]} \alpha\left(h_j^{(l-1)}, \sum_{i \in \mathcal{N}(j)} \psi\left(h_j^{(l-1)}, e_{i,j}, h_i^{(l-1)}\right)\right)\right). \tag{2}$$

- After L iterations compute the graph representation

$$h_X = \tau\left(\frac{1}{|V(X)|} \sum_{k \in V(X)} h_k^{(L)}\right), \tag{3}$$

where τ is a readout function of appropriate dimension.

PENN Class

For $K > 0$ and $L \in \mathbb{N}$ we consider the following subclass of PENNs:

$$\mathcal{H}_{L,K} := \left\{ g : X \mapsto h_X \;\middle|\; \begin{array}{r} h_X \text{ is computed by (2) and (3),} \\ \rho, \alpha, \phi, \psi, \text{ and } \tau \text{ are } K\text{-Lipschitz FNNs,} \\ \phi(\cdot, \mathbf{0}, \cdot) = \psi(\cdot, \mathbf{0}, \cdot) = \mathbf{0}. \end{array} \right\}$$

Lipschitz Continuity

Proposition

Let $X, Y \in \mathcal{G}_N$, and let $g \in \mathcal{H}_{L,K}$ for some $L \in \mathbb{N}$ and $K > 0$. Then for $c(K, L, N) = 4^L 3^{L-1} N^{2(L-1)} K^{3L+1}$,

$$\|g(X) - g(Y)\| \leq c \cdot d_L(X, Y).$$

In the 1-layer case:

Proposition

Let $X, Y \in \mathcal{G}_N$, and let $g \in \mathcal{H}_{1,K}$ for some $K > 0$. Then

$$\|g(X) - g(Y)\| \leq 4K^4 d_1(X, Y).$$

Proof Idea

$$
\begin{aligned}
\|h_k^{X^*} - h_{\bar{k}}^{Y^*}\| \leq K_\rho \Bigg\| & \left(h_k^X, \sum_{i \in \mathcal{N}(k)} \phi(h_k^X, e_{i,k}^X, h_i^X), \sum_{j \in [N]} \alpha \left(h_j^X, \sum_{i \in \mathcal{N}(j)} \psi(h_j^X, e_{i,j}^X, h_i^X) \right) \right) \\
& - \left(h_{\bar{k}}^Y, \sum_{i \in \mathcal{N}(\bar{k})} \phi(h_{\bar{k}}^Y, e_{i,\bar{k}}^Y, h_i^Y), \sum_{j \in [N]} \alpha \left(h_j^Y, \sum_{i \in \mathcal{N}(j)} \psi(h_j^Y, e_{i,j}^Y, h_i^Y) \right) \right) \Bigg\| \\
\leq K_\rho \Bigg[& \underbrace{\left\| h_k^X - h_{\bar{k}}^Y \right\|}_{\leq \|(\cdot,\cdot,h_k^X) - (\cdot,\cdot,h_{\bar{k}}^Y)\| \leq d_L(X,Y)} + \left\| \sum_{i \in \mathcal{N}(k)} \phi(h_k^X, e_{i,k}^X, h_i^X) - \sum_{i \in \mathcal{N}(\bar{k})} \phi(h_{\bar{k}}^Y, e_{i,\bar{k}}^Y, h_i^Y) \right\| \\
& + \left\| \sum_{j \in [N]} \alpha \left(h_j^X, \sum_{i \in \mathcal{N}(j)} \psi(h_j^X, e_{i,j}^X, h_i^X) \right) - \sum_{j \in [N]} \alpha \left(h_j^Y, \sum_{i \in \mathcal{N}(j)} \psi(h_j^Y, e_{i,j}^Y, h_i^Y) \right) \right\| \Bigg] \leq \dots
\end{aligned}
$$

Generalization Bounds

Proposition

Let $\mathcal{G}_N$ be a compact subset of $\mathcal{D}_{N,d,d'}$ and let d_1 denote the pseudo-metric on $\mathcal{G}_N$. Let $\mathcal{Y} \subseteq \{0,1\}$. Let $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ be a loss function bounded by $M_\ell > 0$. Let $K_{\mathcal{H}} > 0$ and let $\mathcal{H}_{1,K_{\mathcal{H}}} \subset \mathcal{X}^{\mathcal{G}_n}$ be the set of 1-layer PENNs defined previously.
Let $k \in \mathbb{N}$, $\mathcal{A} : \mathcal{Z}^k \to \mathcal{H}_{1,K_{\mathcal{H}}}$ be any learning algorithm, and $\mathcal{S}$ be a random sample of size k of $\mathcal{Z}$. Then, for any $\delta, \varepsilon > 0$, with probability at least $1 - \delta$

$$|\ell_{\exp}(\mathcal{A}_s) - \ell_{\text{emp}}(\mathcal{A}_s)| \leq 2\varepsilon + M_\ell \sqrt{\frac{2K \ln(2) + 2\ln(1/\delta)}{k}},$$

where

$$K = 2 \cdot \mathcal{N}(\mathcal{G}_N, d_1, \varepsilon/(8K_{\mathcal{H}}^4)).$$

Reminder: Covering number is the smallest number of ε-balls needed to cover a set.

Covering Number Bound

Connection between the graph pseudo-metric d_L and the metric induced by the Frobenius norm on the space of tensor representations $\mathbb{R}^{N^2 \times (d+d'+d)}$.

Proposition

Let $X, Y \in \mathcal{G}_n$. Let $Z^X \in \mathbb{R}^{n^2 \times (d+d'+d)}$ be a tensor representation of X defined by $Z^X_{i,j,:} = (h^X_j, e^X_{i,j}, h^X_i)$ for $i, j \in [n]$, and define Z^Y analogously. Then

$$d_1(X, Y) \leq 2 \cdot \|Z^X - Z^Y\|_F^2,$$
$$d_L(X, Y) \leq (1 + |V(\mathcal{T}_{L-2}(X))|)\|Z^X - Z^Y\|_F^2, \qquad L \in \mathbb{N}_{\geq 2}.$$

Conclusion

Previous talk:

- We introduce systemic risk measures of **graph valued** financial networks based on **random allocations**; see Gonon et al. [4].
- To compute them we propose **PENNs**.

This talk:

- We introduce *random node features*.
- We show that PENNs with random node features are **universal** approximators of permutation-equivariant functions .
- We also provide **approximation rates**.
- We prove **generalization bounds** for PENNs.

References I

[1] R. Abboud, I. I. Ceylan, M. Grohe, and T. Lukasiewicz. The surprising power of graph neural networks with random node initialization. *arXiv preprint arXiv:2010.01179*, 2020.

[2] M. Bechler-Speicher, M. Eliasof, C.-B. Schönlieb, R. Gilad-Bachrach, and A. Globerson. On the utilization of unique node identifiers in graph neural networks. *arXiv preprint arXiv:2411.02271*, 2024.

[3] G. Dasoulas, L. D. Santos, K. Scaman, and A. Virmaux. Coloring graph neural networks for node disambiguation. *arXiv preprint arXiv:1912.06058*, 2019.

[4] L. Gonon, T. Meyer-Brandis, and N. Weber. Computing systemic risk measures with graph neural networks. *SIAM Journal on Financial Mathematics*, 17(2):565–619, 2026. doi: $10.1137/24\mathrm{M}1697402$. URL `https://doi.org/10.1137/24M1697402`.

[5] N. Keriven and G. Peyré. Universal invariant and equivariant graph neural networks. *Advances in Neural Information Processing Systems*, 32, 2019.

[6] H. Maron, H. Ben-Hamu, N. Shamir, and Y. Lipman. Invariant and equivariant graph networks. *arXiv preprint arXiv:1812.09902*, 2018.

References II

[7] H. Maron, H. Ben-Hamu, H. Serviansky, and Y. Lipman. Provably powerful graph networks. *Advances in neural information processing systems*, 32, 2019.

[8] H. Maron, E. Fetaya, N. Segol, and Y. Lipman. On the universality of invariant networks. In *International conference on machine learning*, pages 4363–4371. PMLR, 2019.

[9] C. Morris, M. Ritzert, M. Fey, W. L. Hamilton, J. E. Lenssen, G. Rattan, and M. Grohe. Weisfeiler and leman go neural: Higher-order graph neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 4602–4609, 2019.

[10] R. Murphy, B. Srinivasan, V. Rao, and B. Ribeiro. Relational pooling for graph representations. In *International Conference on Machine Learning*, pages 4663–4673. PMLR, 2019.

[11] O. Puny, H. Ben-Hamu, and Y. Lipman. Global attention improves graph networks generalization. *arXiv preprint arXiv:2006.07846*, 2020.

References III

[12] R. Sato, M. Yamada, and H. Kashima. Random features strengthen graph neural networks. In *Proceedings of the 2021 SIAM international conference on data mining (SDM)*, pages 333–341. SIAM, 2021.

[13] H. Xu and S. Mannor. Robustness and generalization. *Machine learning*, 86(3):391–423, 2012.

Thank you for your attention!

Questions?