A. Proof of Proposition 8

First, we prove that $\mathcal{M}_{WL}$ is weaker than $\mathcal{M}_{anon}$. It suffices to note that $\mathcal{M}_{WL} \subseteq \mathcal{M}_{anon}$.

It remains to argue that $\mathcal{M}_{anon}$ is weaker than $\mathcal{M}_{WL}$. The proof is an adaptation of the proofs of Lemma 2 in Xu et al. (2019) and Theorem 5 in Morris et al. (2019). Given a labelled graph we show, by induction on the number of rounds of computation, that $\boldsymbol{\ell}^{(t)}_{M_{WL}} \sqsubseteq \boldsymbol{\ell}^{(t)}_{M}$ for all $M \in \mathcal{M}_{anon}$ and every $t \geq 0$.

Clearly, this holds for $t = 0$ since $\boldsymbol{\ell}^{(0)}_{M_{WL}} = \boldsymbol{\ell}^{(0)}_{M} := \boldsymbol{\nu}$, by definition. We assume next that the induction hypothesis holds up to round $t-1$ and consider round t. Let v and w be two vertices such that $(\boldsymbol{\ell}^{(t)}_{M_{WL}})_v = (\boldsymbol{\ell}^{(t)}_{M_{WL}})_w$ holds. This implies, by the definition of M_{WL}, that $(\boldsymbol{\ell}^{(t-1)}_{M_{WL}})_v = (\boldsymbol{\ell}^{(t-1)}_{M_{WL}})_w$ and

$$\{\!\{(\boldsymbol{\ell}^{(t-1)}_{M_{WL}})_u \mid u \in N_G(v)\}\!\} = \{\!\{(\boldsymbol{\ell}^{(t-1)}_{M_{WL}})_u \mid u \in N_G(w)\}\!\}.$$

By the induction hypothesis, this implies that $(\boldsymbol{\ell}^{(t-1)}_{M})_v = (\boldsymbol{\ell}^{(t-1)}_{M})_w$ and

$$\{\!\{(\boldsymbol{\ell}^{(t-1)}_{M})_u \mid u \in N_G(v)\}\!\} = \{\!\{(\boldsymbol{\ell}^{(t-1)}_{M})_u \mid u \in N_G(w)\}\!\}.$$

As a consequence, there is a bijection between $N_G(v)$ and $N_G(w)$ such that to every vertex $u \in N_G(v)$ we can assign a unique vertex $u' \in N_G(w)$ such that $(\boldsymbol{\ell}^{(t-1)}_{M})_u = (\boldsymbol{\ell}^{(t-1)}_{M})_{u'}$. Hence,

$$\textsc{Msg}^{(t)}\left((\boldsymbol{\ell}^{(t-1)}_{M})_v, (\boldsymbol{\ell}^{(t-1)}_{M})_u, -, -\right) = \textsc{Msg}^{(t)}\left((\boldsymbol{\ell}^{(t-1)}_{M})_w, (\boldsymbol{\ell}^{(t-1)}_{M})_{u'}, -, -\right).$$

Since this mapping between $N_G(v)$ and $N_G(w)$ is a bijection we also have:

$$\mathbf{m}^{(t)}_v = \sum_{u \in N_G(v)} \textsc{Msg}^{(t)}\left((\boldsymbol{\ell}^{(t-1)}_{M})_v, (\boldsymbol{\ell}^{(t-1)}_{M})_u, -, -\right) = \sum_{u' \in N_G(w)} \textsc{Msg}^{(t)}\left((\boldsymbol{\ell}^{(t-1)}_{M})_w, (\boldsymbol{\ell}^{(t-1)}_{M})_{u'}, -, -\right) = \mathbf{m}^{(t)}_w.$$

We may thus conclude that

$$(\boldsymbol{\ell}^{(t)}_{M})_v = \textsc{Upd}^{(t)}\left((\boldsymbol{\ell}^{(t-1)}_{M})_v, \mathbf{m}^{(t)}_v\right) = \textsc{Upd}^{(t)}\left((\boldsymbol{\ell}^{(t-1)}_{M})_w, \mathbf{m}^{(t)}_w\right) = (\boldsymbol{\ell}^{(t)}_{M})_w,$$

as desired. □

B. Anonymous MPNNs and self labels

We connect to MPNNs of the form $f^{(t)}_{comb}(\boldsymbol{\ell}^{(t-1)}_v, f^{(t)}_{aggr}(\{\!\{\boldsymbol{\ell}^{(t-1)}_u \mid u \in N_G(v)\}\!\}))$ used in Xu et al. (2019) and Morris et al. (2019). Observe that the aggregation functions $f^{(t)}_{aggr}(\{\!\{\boldsymbol{\ell}^{(t-1)}_u \mid u \in N_G(v)\}\!\})$ can be written in the form $g^{(t)}(\sum_{u \in N_G(v)} h^{(t)}(\boldsymbol{\ell}^{(t-1)}_u))$, based on Lemma 5 from Xu et al. (2019).

Suppose that $\boldsymbol{\nu} : V \to \mathbb{A}^{s_0}$. It suffices to define for every $t \geq 1$, every $\mathbf{x}, \mathbf{y} \in \mathbb{A}^{s_{t-1}}$, every $v \in V$ and $u \in N_G(u)$:

$$\begin{aligned} \textsc{Msg}^{(t)}(\mathbf{x}, \mathbf{y}, -, -) &:= h^{(t)}(\mathbf{y}), \\ \textsc{Upd}^{(t)}(\mathbf{x}, \mathbf{y}) &:= f^{(t)}_{comb}\left(\mathbf{x}, g^{(t)}(\mathbf{y})\right). \end{aligned} \tag{8}$$

Hence, the aMPNNs in Xu et al. (2019) and Morris et al. (2019) are examples of our aMPNNs.

The aMPNNs that we consider in this paper are slightly more general than those defined by (8). Indeed, we consider message functions that can also depend on the previous label $\boldsymbol{\ell}^{(t-1)}_v$. In contrast, the message functions in (8) only depend on $\mathbf{y}$, which corresponds to the previous labels $\boldsymbol{\ell}^{(t-1)}_u$ of neighbours $u \in N_G(v)$. Let $\mathcal{M}^{-}_{anon}$ denote the class of aMPNNs whose message functions only depend on the previous labels of neighbours. It now suffices to observe (see Example 2) that $M_{WL} \in \mathcal{M}^{-}_{anon}$ to infer, combined with Proposition 8, that:

Corollary 20. *The classes $\mathcal{M}^{-}_{anon}$, $\mathcal{M}_{anon}$ and $\mathcal{M}_{WL}$ are all equally strong.*

We observe, however, that this does not imply that for every aMPNN M in $\mathcal{M}_{anon}$ there exists an aMPNN M' in $\mathcal{M}^{-}_{anon}$ such that $\boldsymbol{\ell}^{(t)}_{M} \equiv \boldsymbol{\ell}^{(t)}_{M'}$ for all $t \geq 0$. Indeed, the corollary implies that for every M in $\mathcal{M}_{anon}$ there exists an aMPNN M' in $\mathcal{M}^{-}_{anon}$ such that $M \preceq M'$, and there exists an M'' in $\mathcal{M}_{anon}$, possibly different from M, such that $M' \preceq M''$. In fact, such an aMPNN M'', in this case is M_{WL}.

C. Proof of Theorem 11

Crucial in the proof is the notion of row-independence modulo equality from Morris et al. (2019), which we recall next.

Definition 21 (Row-independence modulo equality)**.** A labelling $\boldsymbol{\ell} : V \to \mathbb{A}^s$ is *row-independent modulo equality* if the set of unique labels assigned by $\boldsymbol{\ell}$ is linearly independent.

In what follows, we always assume that the initial labelling $\boldsymbol{\nu}$ of G is row-independent modulo equality. One can always ensure this by extending the labels.[7]

Proof. We already know that $\mathcal{M}_{\mathit{GNN}}^{\mathit{ReLU}}$ is weaker than $\mathcal{M}_{\mathit{WL}}$ (Theorem 10 and also Corollary 9). It remains to show that $\mathcal{M}_{\mathit{WL}}$ is weaker than $\mathcal{M}_{\mathit{GNN}}^{\mathit{ReLU}}$. That is, given an aMPNN M_{WL}, we need to construct an aMPNN M in $\mathcal{M}_{\mathit{GNN}}^{\mathit{ReLU}}$ such that $\boldsymbol{\ell}_M^{(t)} \sqsubseteq \boldsymbol{\ell}_{M_{\mathrm{WL}}}^{(t)}$, for all $t \geq 0$. We observe that since $\boldsymbol{\ell}_{M_{\mathrm{WL}}}^{(t)} \sqsubseteq \boldsymbol{\ell}_M^{(t)}$ for any M in $\mathcal{M}_{\mathit{GNN}}^{\mathit{ReLU}}$, this is equivalent to constructing an M such that $\boldsymbol{\ell}_M^{(t)} \equiv \boldsymbol{\ell}_{M_{\mathrm{WL}}}^{(t)}$.

The proof is by induction on the number of computation rounds. The aMPNN M in $\mathcal{M}_{\mathit{GNN}}^{\mathit{ReLU}}$ that we will construct will use message and update functions of the form:

$$\textsc{Msg}^{(t)}(\mathbf{x}, \mathbf{y}, -, -) := \mathbf{y}\mathbf{W}^{(t)} \text{ and } \textsc{Upd}^{(t)}(\mathbf{x}, \mathbf{y}) := \mathrm{ReLU}\left(p\mathbf{x}\mathbf{W}^{(t)} + \mathbf{y} + \mathbf{b}^{(t)}\right) \tag{9}$$

for some value $p \in \mathbb{A}$, $0 < p < 1$, weight matrix $\mathbf{W}^{(t)} \in \mathbb{A}^{s_{t-1} \times s_t}$, and bias vector $\mathbf{b}^{(t)} \in \mathbb{A}^{s_t}$. Note that, in contrast to aMPNNs of the form (2), we only have one weight matrix per round, instead of two, at the cost of introducing an extra parameter $p \in \mathbb{A}$. Furthermore, the aMPNN constructed in Morris et al. (2019) uses two distinct weight matrices in $\mathbb{A}^{(s_{t-1}+s_0)\times(s_t+s_0)}$ (we come back to this at the end of this section) whereas our weight matrices are elements of $\mathbb{A}^{s_{t-1}\times s_t}$ and thus of smaller dimension.

The induction hypothesis is that $\boldsymbol{\ell}_M^{(t)} \equiv \boldsymbol{\ell}_{M_{\mathrm{WL}}}^{(t)}$ and that $\boldsymbol{\ell}_M^{(t)}$ is row-independent modulo equality.

For $t = 0$, we have that for any $M \in \mathcal{M}_{\mathit{GNN}}^{\mathit{ReLU}}$, $\boldsymbol{\ell}_M^{(0)} = \boldsymbol{\ell}_{M_{\mathrm{WL}}}^{(0)} := \boldsymbol{\nu}$, by definition. Moreover, $\boldsymbol{\ell}_M^{(0)}$ is row-independent modulo equality because $\boldsymbol{\nu}$ is so, by assumption.

We next assume that up to round $t-1$ we have found weight matrices and bias vectors for M such that $\boldsymbol{\ell}_M^{(t-1)}$ satisfies the induction hypothesis. We will show that for round t we can find a weight matrix $\mathbf{W}^{(t)} \in \mathbb{A}^{s_{t-1}\times s_t}$ and bias vector $\mathbf{b}^{(t)} \in \mathbb{A}^{s_t}$ such that $\boldsymbol{\ell}_M^{(t)}$ also satisfies the hypothesis.

Let $\mathbf{L}^{(t-1)} \in \mathbb{A}^{n \times s_{t-1}}$ denote the matrix consisting of rows $(\boldsymbol{\ell}_M^{(t-1)})_v$, for $v \in V$. Moreover, we denote by $\mathsf{uniq}(\mathbf{L}^{(t-1)})$ a $(m \times s_{t-1})$-matrix consisting of the m unique rows in $\mathbf{L}^{(t-1)}$ (the order of rows is irrelevant). We denote the rows in $\mathsf{uniq}(\mathbf{L}^{(t-1)})$ by $\mathbf{a}_1, \ldots, \mathbf{a}_m \in \mathbb{A}^{s_{t-1}}$. By the induction hypothesis, these rows are linearly independent. Following the same argument as in Morris et al. (2019), this implies that there exists an $(s_{t-1} \times m)$-matrix $\mathbf{U}^{(t)}$ such that $\mathsf{uniq}(\mathbf{L}^{(t-1)})\mathbf{U}^{(t)} = \mathbf{I}$. Let us denote by $\mathbf{e}_1, \ldots, \mathbf{e}_m \in \mathbb{A}^m$ the rows of $\mathbf{I}$. In other words, in $\mathbf{e}_i$, all entries are zero except for entry i, which holds value 1.

We consider the following intermediate labelling $\boldsymbol{\mu}^{(t)} : V \to \mathbb{A}^m$ defined by

$$v \mapsto \left((\mathbf{A} + p\mathbf{I})\mathbf{L}^{(t-1)}\mathbf{U}^{(t)}\right)_v. \tag{10}$$

We know that for every vertex v, $(\boldsymbol{\ell}_M^{(t-1)})_v$ corresponds to a unique row $\mathbf{a}_i$ in $\mathsf{uniq}(\mathbf{L}^{(t-1)})$. We denote the index of this row by $\rho(v)$. More specifically, $(\boldsymbol{\ell}_M^{(t-1)})_v = \mathbf{a}_{\rho(v)}$. Let $N_G(v,i) := \{u \mid u \in N_G(v), \rho(v) = i\}$. That is, $N_G(v,i)$ consists of all neighbours u of v which are labelled as $\mathbf{a}_i$ by $\boldsymbol{\ell}_M^{(t-1)}$. It is now readily verified that the label $\boldsymbol{\mu}_v^{(t)}$ defined in (10) is of the form

$$\boldsymbol{\mu}_v^{(t)} = p\mathbf{e}_{\rho(v)} + \sum_{i=1}^{m} |N_G(v,i)|\mathbf{e}_i. \tag{11}$$

We clearly have that $\boldsymbol{\ell}_{M_{\mathrm{WL}}}^{(t)} \sqsubseteq \boldsymbol{\mu}^{(t)}$. The converse also holds, as is shown in the following lemma.

[7] Note that we allow to extend the labels only in a restricted way that does not impact the initial labelling, namely: equal labels must be extended in the same way.

Lemma 22. *For any two vertices v and w, we have that $\boldsymbol{\mu}_v^{(t)} = \boldsymbol{\mu}_w^{(t)}$ implies $(\boldsymbol{\ell}_{M_{\mathsf{WL}}}^{(t)})_v = (\boldsymbol{\ell}_{M_{\mathsf{WL}}}^{(t)})_w$.*

Proof. We argue by contradiction. Suppose, for the sake of contradiction, that there exist two vertices $v, w \in V$ such that

$$\boldsymbol{\mu}_v^{(t)} = \boldsymbol{\mu}_w^{(t)} \text{ and } (\boldsymbol{\ell}_{M_{\mathsf{WL}}}^{(t)})_v \neq (\boldsymbol{\ell}_{M_{\mathsf{WL}}}^{(t)})_w \tag{12}$$

hold. We show that this is impossible for any value p satisfying $0 < p < 1$. (Recall from (11) that $\boldsymbol{\mu}_v^{(t)}$ depends on p.)

We distinguish between the following two cases. If $(\boldsymbol{\ell}_{M_{\mathsf{WL}}}^{(t)})_v \neq (\boldsymbol{\ell}_{M_{\mathsf{WL}}}^{(t)})_w$ then either

(i) $(\boldsymbol{\ell}_{M_{\mathsf{WL}}}^{(t-1)})_v \neq (\boldsymbol{\ell}_{M_{\mathsf{WL}}}^{(t-1)})_w$; or

(ii) $(\boldsymbol{\ell}_{M_{\mathsf{WL}}}^{(t-1)})_v = (\boldsymbol{\ell}_{M_{\mathsf{WL}}}^{(t-1)})_w$ and $\{\!\{(\boldsymbol{\ell}_{M_{\mathsf{WL}}}^{(t-1)})_u \mid u \in N_G(v)\}\!\} \neq \{\!\{(\boldsymbol{\ell}_{M_{\mathsf{WL}}}^{(t-1)})_u \mid u \in N_G(w)\}\!\}$.

We first consider case (i). Observe that $(\boldsymbol{\ell}_{M_{\mathsf{WL}}}^{(t-1)})_v \neq (\boldsymbol{\ell}_{M_{\mathsf{WL}}}^{(t-1)})_w$ implies that $(\boldsymbol{\ell}_{M}^{(t-1)})_v \neq (\boldsymbol{\ell}_{M}^{(t-1)})_w$. This follows from the induction hypothesis $\boldsymbol{\ell}_{M}^{(t-1)} \equiv \boldsymbol{\ell}_{M_{\mathsf{WL}}}^{(t-1)}$. It now suffices to observe that $\boldsymbol{\mu}_v^{(t)} = \boldsymbol{\mu}_w^{(t)}$ implies that the corresponding linear combinations, as described in (11), satisfy:

$$p\mathbf{e}_{\rho(v)} + \sum_{i=1}^{m} |N_G(v,i)|\mathbf{e}_i = p\mathbf{e}_{\rho(w)} + \sum_{i=1}^{m} |N_G(w,i)|\mathbf{e}_i.$$

We can assume, without loss of generality, that $(\boldsymbol{\ell}_{M}^{(t-1)})_v = \mathbf{a}_1$ and $(\boldsymbol{\ell}_{M}^{(t-1)})_w = \mathbf{a}_2$. Recall that $\mathbf{a}_1$ and $\mathbf{a}_2$ are two distinct labels. Then, the previous equality implies:

$$(|N_G(v,1)|+p - |N_G(w,1)|)\,\mathbf{e}_1 + (|N_G(v,2)|-|N_G(w,2)|-p)\,\mathbf{e}_2 + \sum_{i=3}^{m} (|N_G(v,i)|-|N_G(w,i)|)\,\mathbf{e}_i = 0.$$

Since $\mathbf{e}_1, \dots, \mathbf{e}_m$ are linearly independent, this implies that $|N_G(v,i)|-|N_G(w,i)|= 0$ for all $i = 3, \dots, m$ and $|N_G(v,1)|+p - |N_G(w,1)|= 0$ and $|N_G(v,2)|-|N_G(w,2)|-p = 0$. Since $|N_G(v,1)|-|N_G(w,1)|\in \mathbb{Z}$ and $0 < p < 1$, this is impossible. We may thus conclude that case (i) cannot occur.

Suppose next that we are in case (ii). Recall that for case (ii), we have that $(\boldsymbol{\ell}_{M_{\mathsf{WL}}}^{(t-1)})_v = (\boldsymbol{\ell}_{M_{\mathsf{WL}}}^{(t-1)})_w$ and thus also $(\boldsymbol{\ell}_{M}^{(t-1)})_v = (\boldsymbol{\ell}_{M}^{(t-1)})_w$. Using the same notation as above, we may assume that $(\boldsymbol{\ell}_{M}^{(t-1)})_v = (\boldsymbol{\ell}_{M}^{(t-1)})_w = \mathbf{a}_1$. In case (ii), however, we have that $\{\!\{(\boldsymbol{\ell}_{M_{\mathsf{WL}}}^{(t-1)})_u \mid u \in N_G(v)\}\!\} \neq \{\!\{(\boldsymbol{\ell}_{M_{\mathsf{WL}}}^{(t-1)})_u \mid u \in N_G(w)\}\!\}$ and thus also $\{\!\{(\boldsymbol{\ell}_{M}^{(t-1)})_u \mid u \in N_G(v)\}\!\} \neq \{\!\{(\boldsymbol{\ell}_{M}^{(t-1)})_u \mid u \in N_G(w)\}\!\}$. That is, there must exist a label assigned by $\boldsymbol{\ell}_{M}^{(t-1)}$ that does not occur the same number of times in the neighbourhoods of v and w, respectively. Suppose that this label is $\mathbf{a}_2$. The case when this label is $\mathbf{a}_1$ can be treated similarly. It now suffices to observe that $\boldsymbol{\mu}_v^{(t)} = \boldsymbol{\mu}_w^{(t)}$ implies that the corresponding linear combinations, as described in (11), satisfy:

$$(|N_G(v,1)|+p)\,\mathbf{e}_1 + |N_G(v,2)|\mathbf{e}_2 + \sum_{i=3}^{m} |N_G(v,i)|\mathbf{e}_i = (|N_G(w,1)|+p)\,\mathbf{e}_1 + |N_G(w,2)|\mathbf{e}_2 + \sum_{i=3}^{m} |N_G(w,i)|\mathbf{e}_i.$$

Using a similar argument as before, based on the linear independence of $\mathbf{e}_1, \dots, \mathbf{e}_m$, we can infer that $|N_G(v,2)|= |N_G(w,2)|$. We note, however, that $\mathbf{a}_2$ appeared a different number of times among the neighbours of v and w. Hence, also case (ii) is ruled out and our assumption (12) is invalid. This implies $\boldsymbol{\mu}^{(t)} \sqsubseteq \boldsymbol{\ell}_{M_{\mathsf{WL}}}^{(t)}$, as desired and thus concludes the proof of the lemma. □

From here, to continue with the proof of Theorem 11, we still need to take care of the ReLU activation function. Importantly, its application should ensure row-independence modulo equality and make sure the labelling "refines" $\boldsymbol{\ell}_{M_{\mathsf{WL}}}^{(t)}$. To do so, we again follow closely the proof strategy of Morris et al. (2019). More specifically, we will need an analogue of the following result. In the sequel we denote by $\mathbf{J}$ a matrix with all entries having value 1 and whose size will be determined from the context.

Lemma 23 (Lemma 9 from Morris et al., 2019)**.** *Let $\mathbf{C} \in \mathbb{A}^{m\times w}$ be a matrix in which all entries are non-negative and all rows are pairwise disjoint. Then there exists a matrix $\mathbf{X} \in \mathbb{A}^{w\times m}$ such that* $\mathrm{sign}(\mathbf{CX} - \mathbf{J})$ *is a non-singular matrix in $\mathbb{A}^{m\times m}$.*

We prove the following for the ReLU function.

Lemma 24. *Let $\mathbf{C} \in \mathbb{A}^{m\times w}$ be a matrix in which all entries are non-negative, all rows are pairwise disjoint and such that no row consists entirely out of zeroes*[8]. *Then there exists a matrix $\mathbf{X} \in \mathbb{A}^{w\times m}$ and a constant $q \in \mathbb{A}$ such that $\mathrm{ReLU}(\mathbf{CX} - q\mathbf{J})$ is a non-singular matrix in $\mathbb{A}^{m\times m}$.*

Proof. Let C be the maximal entry in $\mathbf{C}$ and consider the column vector $\mathbf{z} = (1, C, C^2, \dots, C^{w-1})^{\mathsf{T}} \in \mathbb{A}^{w\times 1}$. Then each entry in $\mathbf{c} = \mathbf{Cz} \in \mathbb{A}^{m\times 1}$ is positive and all entries in $\mathbf{c}$ are pairwise distinct. Let $\mathbf{P}$ be a permutation matrix in $\mathbb{A}^{m\times m}$ such that $\mathbf{c}' = \mathbf{Pc}$ is such that $\mathbf{c}' = (c'_1, c'_2, \dots, c'_m)^{\mathsf{T}} \in \mathbb{A}^{m\times 1}$ with $c'_1 > c'_2 > \cdots > c'_m > 0$. Consider $\mathbf{x} = \left(\frac{1}{c'_1}, \dots, \frac{1}{c'_m}\right) \in \mathbb{A}^{1\times m}$. Then, for $\mathbf{E} = \mathbf{c}'\mathbf{x} \in \mathbb{A}^{m\times m}$

$$\mathbf{E}_{ij} = \frac{c'_i}{c'_j} \text{ and } \mathbf{E}_{ij} = \begin{cases} 1 & \text{if } i = j \\ > 1 & \text{if } i < j \\ < 1 & \text{if } i > j. \end{cases}$$

Let q be the greatest value in $\mathbf{E}$ smaller than 1. Consider $\mathbf{F} = \mathbf{E} - q\mathbf{J}$. Then,

$$\mathbf{F}_{ij} = \frac{c'_i}{c'_j} - q \text{ and } \mathbf{F}_{ij} = \begin{cases} 1-q & \text{if } i = j \\ > 0 & \text{if } i < j \\ \leq 0 & \text{if } i > j. \end{cases}$$

As a consequence,

$$\mathrm{ReLU}(\mathbf{F})_{ij} = \begin{cases} 1-q & \text{if } i = j \\ > 0 & \text{if } i < j \\ 0 & \text{if } i > j. \end{cases}$$

This is an upper triangular matrix with (nonzero) value $1-q$ on its diagonal. It is therefore non-singular.

We now observe that $\mathbf{Q}\mathrm{ReLU}(\mathbf{F}) = \mathrm{ReLU}(\mathbf{QF})$ for any row permutation $\mathbf{Q}$. Furthermore, non-singularity is preserved under row permutations and $\mathbf{QJ} = \mathbf{J}$. Hence, if we define $\mathbf{X} = \mathbf{zx}$ and use the permutation matrix $\mathbf{P}$, then:

$$\mathbf{P}\mathrm{ReLU}(\mathbf{CX} - q\mathbf{J}) = \mathrm{ReLU}(\mathbf{PCzx} - q\mathbf{PJ}) = \mathrm{ReLU}(\mathbf{E} - q\mathbf{J}) = \mathrm{ReLU}(\mathbf{F}),$$

and we have that $\mathrm{ReLU}(\mathbf{CX} - q\mathbf{J})$ is non-singular, as desired. This concludes the proof of the lemma. □

We now apply this lemma to the matrix $\mathsf{uniq}(\mathbf{M}^{(t)})$, with $\mathbf{M}^{(t)} \in \mathbb{A}^{n\times m}$ consisting of the rows $\boldsymbol{\mu}_v^{(t)}$, for $v \in V$. Inspecting the expression from Equation (11) for $\boldsymbol{\mu}_v^{(t)}$ we see that each row in $\mathbf{M}^{(t)}$ holds non-negative values and no row consists entirely out of zeroes. Let $\mathbf{X}^{(t)}$ and $q^{(t)}$ be the matrix and constant returned by Lemma 24 such that $\mathrm{ReLU}\left(\mathsf{uniq}(\mathbf{M}^{(t)})\mathbf{X}^{(t)} - q^{(t)}\mathbf{J}\right)$ is an $m \times m$ non-singular matrix. We now define

$$\boldsymbol{\ell}_M^{(t)} := \mathrm{ReLU}\left(\mathbf{M}^{(t)}\mathbf{X}^{(t)} - q^{(t)}\mathbf{J}\right).$$

From the non-singularity of $\mathrm{ReLU}\left(\mathsf{uniq}(\mathbf{M}^{(t)})\mathbf{X}^{(t)} - q^{(t)}\mathbf{J}\right)$ we can immediately infer that $\boldsymbol{\ell}_M^{(t)}$ is row-independent modulo equality. It remains to argue that $\boldsymbol{\ell}_M^{(t)} \equiv \boldsymbol{\ell}_{M_{\mathrm{WL}}}^{(t)}$. This now follows from the fact that $\boldsymbol{\mu}^{(t)} \equiv \boldsymbol{\ell}_{M_{\mathrm{WL}}}^{(t)}$ and each of the m unique labels assigned by $\boldsymbol{\mu}^{(t)}$ uniquely corresponds to a row in $\mathsf{uniq}(\mathbf{M}^{(t)})$, which in turn can be mapped bijectively to a row in $\mathrm{ReLU}\left(\mathsf{uniq}(\mathbf{M}^{(t)})\mathbf{X}^{(t)} - q^{(t)}\mathbf{J}\right)$. We conclude by observing that the desired weight matrices and bias vector at round t for M are now given by $\mathbf{W}^{(t)} := \mathbf{U}^{(t)}\mathbf{X}^{(t)}$ and $\mathbf{b}^{(t)} := -q^{(t)}\mathbf{1}$. This concludes the proof of Theorem 11. □

[8] Compared to Lemma 23, we additionally require non-zero rows.

D. Refining Theorem 11 for the sign activation function

We remark that the proof of Theorem 11 can be used for $\mathcal{M}^{sign}_{GNN}$ as well. One just has to use Lemma 23 instead of Lemma 24. It is interesting to note that the bias vector for the sign activation function in Lemma 23 is the same for every t. A similar statement holds for the ReLU function. Indeed, we recall that we apply Lemma 24 to $\mathsf{uniq}(\mathbf{M}^{(t)})$. For every t, the entries in this matrix are of the form $i + p$ (which is smaller than $i + 1$) or i, for $i \in \{1, 2, \dots, n\}$. Hence, for every t, the maximal entry (denoted by C in the proof of Lemma 24) is upper bounded by $n + 1$. The value $q^{(t)}$ relates to the largest possible ratios, smaller than 1, of elements in the matrix constructed in Lemma 24.

When the lemma is applied to an $m \times w$ matrix, this ratio is upper bounded by $\frac{(n+1)^w-1}{(n+1)^w}$. Note that, since the lemma is applied to matrices arising from $\boldsymbol{\mu}^{(t)}$, w will always be at most n. Hence, taking any $q^{(t)} := q$ for $\frac{(n+1)^n-1}{(n+1)^n} < q < 1$ suffices. We can take q to be arbitrarily close to 1, but not 1 itself.

We can thus strengthen Theorem 10, as follows. We denote by $\mathcal{M}_{GNN^-}$ the class of aMPNNs using message and update functions of the form:

$$\textsc{Msg}^{(t)}(\mathbf{x}, \mathbf{y}, -, -) := \mathbf{y}\mathbf{W}^{(t)} \text{ and } \textsc{Upd}^{(t)}(\mathbf{x}, \mathbf{y}) := \sigma\left(p\mathbf{x}\mathbf{W}^{(t)} + \mathbf{y} - q\mathbf{1}\right), \tag{13}$$

parameterised with values $p, q \in \mathbb{A}$, $0 \leq p, q \leq 1$ and weight matrices $\mathbf{W}^{(t)} \in \mathbb{A}^{s_{t-1} \times s_t}$, and where σ can be either the sign or ReLU function.

Corollary 25. *The class $\mathcal{M}_{GNN^-}$ is equally strong as $\mathcal{M}_{GNN}$ and is equally strong as $\mathcal{M}_{WL}$.*

E. Proof of Proposition 13

To prove the first part of the claim notice that $\mathcal{M}_{anon}$ is weaker than $\mathcal{M}_{deg}$, simply because any aMPNN is a dMPNN. Then the result follows from Theorem 7.

For the second part it suffices to provide a dMPNN M and a labelled graph $(G, \boldsymbol{\nu})$ such that there exists a round $t \geq 0$ for which $\boldsymbol{\ell}^{(t)}_{M_{WL}} \not\sqsubseteq \boldsymbol{\ell}^{(t)}_{M}$ holds. We construct such an M originating from a GCN (Kipf & Welling, 2017) defined in Example 3. That is, M is a dMPNN in $\mathcal{M}_{dGNN_4}$. Consider the labelled graph $(G, \boldsymbol{\nu})$ with vertex labelling $\boldsymbol{\nu}_{v_1} = \boldsymbol{\nu}_{v_2} = (1, 0, 0)$, $\boldsymbol{\nu}_{v_3} = \boldsymbol{\nu}_{v_6} = (0, 1, 0)$ and $\boldsymbol{\nu}_{v_4} = \boldsymbol{\nu}_{v_5} = (0, 0, 1)$, and edges $\{v_1, v_3\}$, $\{v_2, v_3\}$, $\{v_3, v_4\}$, $\{v_4, v_5\}$, and $\{v_5, v_6\}$, as depicted in Figure 3.

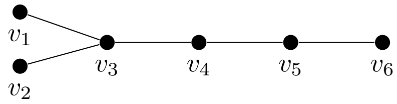

Figure 3. Graph G.

Recall that $\boldsymbol{\ell}^{(0)} = \boldsymbol{\nu}$ and

$$(\boldsymbol{\ell}^{(1)}_M)_v := \text{ReLU}\left(\left(\frac{1}{1+d_v}\right)\boldsymbol{\ell}^{(0)}_v\mathbf{W}^{(1)} + \sum_{u \in N_G(v)}\left(\frac{1}{\sqrt{1+d_v}}\right)\left(\frac{1}{\sqrt{1+d_u}}\right)\boldsymbol{\ell}^{(0)}_u\mathbf{W}^{(1)}\right).$$

We next define $\mathbf{W}^{(1)} := \left(\begin{smallmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{smallmatrix}\right)$. It can be verified that

$$\boldsymbol{\ell}_M^{(1)} = \begin{pmatrix} \frac{1}{2} & 0 & 0 \\ \frac{1}{2} & 0 & 0 \\ 0 & \frac{1}{4} & 0 \\ 0 & 0 & \frac{1}{3} \\ 0 & 0 & \frac{1}{3} \\ 0 & \frac{1}{2} & 0 \end{pmatrix} + \begin{pmatrix} 0 & \frac{1}{2\sqrt{2}} & 0 \\ 0 & \frac{1}{2\sqrt{2}} & 0 \\ \frac{1}{\sqrt{2}} & 0 & \frac{1}{2\sqrt{3}} \\ 0 & \frac{1}{2\sqrt{3}} & \frac{1}{3} \\ 0 & \frac{1}{\sqrt{6}} & \frac{1}{3} \\ 0 & 0 & \frac{1}{\sqrt{6}} \end{pmatrix} = \begin{pmatrix} \frac{1}{2} & \frac{1}{2\sqrt{2}} & 0 \\ \frac{1}{2} & \frac{1}{2\sqrt{2}} & 0 \\ \frac{1}{\sqrt{2}} & \frac{1}{4} & \frac{1}{2\sqrt{3}} \\ 0 & \frac{1}{2\sqrt{3}} & \frac{2}{3} \\ 0 & \frac{1}{\sqrt{6}} & \frac{2}{3} \\ 0 & \frac{1}{2} & \frac{1}{\sqrt{6}} \end{pmatrix}.$$

We observe that $(\boldsymbol{\ell}_M^{(1)})_{v_4} \neq (\boldsymbol{\ell}_M^{(1)})_{v_5}$. We note, however, that $(\boldsymbol{\ell}_{M_{WL}}^{(1)})_{v_4} = \textsc{Hash}((0,0,1), \{\!\{(0,0,1),(0,1,0)\}\!\}) = (\boldsymbol{\ell}_{M_{WL}}^{(1)})_{v_5}$. Hence, $\boldsymbol{\ell}_{M_{WL}}^{(1)} \not\sqsubseteq \boldsymbol{\ell}_M^{(1)}$. □

F. Proof of Lemma 15

We define the aMPNN M_d with the following message and update functions. For each $\mathbf{x}, \mathbf{y} \in \mathbb{A}^s$, $z \in \mathbb{A}$, and vertices $v, u \in N_G(v)$ we define:

$$\textsc{Msg}^{(1)}(\mathbf{x}, \mathbf{y}, -, -) := 1 \text{ and } \textsc{Upd}^{(1)}(\mathbf{x}, z) := (\mathbf{x}, z)\,.$$

Then, $\mathbf{m}_v^{(1)} := \sum_{u \in N_G(v)} 1 = d_v$ and $(\boldsymbol{\ell}_{M_d}^{(1)})_v := \textsc{Upd}^{(1)}(\boldsymbol{\nu}_v, d_v) = (\boldsymbol{\nu}_v, d_v) \in \mathbb{A}^{s+1}$, as desired. □

G. Proof of Proposition 14

By Theorem 7 it suffices to prove that the class $\mathcal{M}_{\mathit{deg}}$ is weaker than $\mathcal{M}_{\mathit{anon}}$, with 1 step ahead. Let $(G, \boldsymbol{\nu})$ be a labelled graph with $\boldsymbol{\nu} : V \to \mathbb{A}^{s_0}$. Take an arbitrary dMPNN M_1 such that for every round $t \geq 1$ the message function is

$$\textsc{Msg}_{M_1}^{(t)}(\mathbf{x}, \mathbf{y}, d_v, d_u) \in \mathbb{A}^{s'_t}$$

and $\textsc{Upd}_{M_1}^{(t)}(\mathbf{x}, \mathbf{z})$ is the update function.

We construct an aMPNN M_2 such that $\boldsymbol{\ell}_{M_2}^{(t+1)} \sqsubseteq \boldsymbol{\ell}_{M_1}^{(t)}$ holds, as follows. We denote the message and update functions of M_2 by $\textsc{Msg}_{M_2}^{(t)}$ and $\textsc{Upd}_{M_2}^{(t)}$, respectively. We will keep as an invariant **(I1)** stating that for all v if we have $\mathbf{x}' = (\boldsymbol{\ell}_{M_1}^{(t-1)})_v \in \mathbb{A}^{s_{t-1}}$ then $\mathbf{x} = (\mathbf{x}', d_v) = (\boldsymbol{\ell}_{M_2}^{(t)})_v \in \mathbb{A}^{s_{t-1}+1}$.

For $t = 1$, we let $\textsc{Msg}_{M_2}^{(1)}$ and $\textsc{Upd}_{M_2}^{(1)}$ be the functions defined by Lemma 15. As a consequence, $(\boldsymbol{\ell}_{M_2}^{(1)})_v = (\boldsymbol{\nu}_v, d_v) \in \mathbb{A}^{s_0+1}$ for every vertex v. We clearly have that $\boldsymbol{\ell}_{M_2}^{(1)} \sqsubseteq \boldsymbol{\ell}_{M_1}^{(0)}$ and the invariant **(I1)** trivially holds.

For $t \geq 2$, we define the message and update functions of M_2 as follows:

$$\textsc{Msg}_{M_2}^{(t)}(\mathbf{x}, \mathbf{y}, -, -) := \textsc{Msg}_{M_1}^{(t-1)}(\mathbf{x}', \mathbf{y}', x, y)$$

where $\mathbf{x} = (\mathbf{x}', x)$ and $\mathbf{y} = (\mathbf{y}', y)$ and by invariant **(I1)** $x = d_v$ and $y = d_u$. Notice that the message function remains anonymous as d_u and d_v are not obtained by setting $f(v) = d_v$ and $f(u) = d_u$ but instead were computed once by the first message aggregation and encoded in the labels of v and u. The update function is defined as follows:

$$\textsc{Upd}_{M_2}^{(t)}(\mathbf{x}, \mathbf{z}) := \left(\textsc{Upd}_{M_1}^{(t-1)}(\mathbf{x}', \mathbf{z}'), x\right) \in \mathbb{A}^{s_{t-1}+1},$$

where $\mathbf{x} = (\mathbf{x}', x)$ and by invariant **(I1)** $x = d_v$. In other words, in each round $t \geq 2$, M_2 extracts the degrees from the last entries in the labels and simulates round $t-1$ of M_1. It is readily verified that $\boldsymbol{\ell}_{M_2}^{(t)} \sqsubseteq \boldsymbol{\ell}_{M_1}^{(t-1)}$ for every t, as desired and that the invariant **(I1)** holds. □

H. GCNs as dGNNs of the form (5)

Example 26. The GCN architecture of Kipf & Welling (2017) corresponds to graph neural networks of the form (5), with $\mathbf{W}_1^{(t)} = \mathbf{0} \in \mathbb{A}^{s_{t-1} \times s_t}$, $p = 1$, $\mathbf{b}^{(t)} = \mathbf{0} \in \mathbb{A}^{s}$, and where $\mathbf{g} = \mathbf{h}$ are defined by the function $g(n) = h(n) = (1+n)^{-1/2}$. □

I. Proof of Proposition 16

We first show a more general result, related to graph neural networks of the form (5) in which $\mathsf{diag}(\mathbf{h}) = \mathbf{I}$. In other words, the function $h : \mathbb{N}^{+} \to \mathbb{A}$ underlying $\mathbf{h}$ is the constant one function, i.e., $h(n) = 1$ for all $n \in \mathbb{N}^{+}$.

Proposition 27. *The subclass of $\mathcal{M}_{\mathit{dGNN}}$, in which the function h is the constant one function, is weaker than $\mathcal{M}_{\mathit{WL}}$.*

Proof. We show that any MPNN M in this class is an anonymous MPNNs. To see this, it suffices to observe that any dMPNN in $\mathcal{M}_{\mathit{dGNN}}$, and thus also M in particular, is equivalent to a dMPNN with message and update functions defined as follows. For every round $t \geq 1$, every $\mathbf{x}, \mathbf{y} \in \mathbb{A}^{s_{t-1}}$, $\mathbf{z} = (\mathbf{z}', z) \in \mathbb{A}^{s_t + 1}$, and every vertex v and $u \in N_G(v)$:

$$\textsc{Msg}^{(t)}(\mathbf{x}, \mathbf{y}, d_v, d_u) := \left(h(d_u)\mathbf{y}\mathbf{W}_2^{(t)}, 1\right) \in \mathbb{A}^{s_t + 1} \tag{14}$$

and

$$\textsc{Upd}^{(t)}(\mathbf{x}, \mathbf{z}) := \sigma\left(\mathbf{x}\mathbf{W}_1^{(t)} + g(z)\mathbf{z}' + pg(z)h(z)\mathbf{x}\mathbf{W}_2^{(t)} + \mathbf{b}^{(t)}\right) \in \mathbb{A}^{s_t}, \tag{15}$$

where $z \in \mathbb{A}$ will hold the degree information of the vertex under consideration (i.e., d_v) after message passing. That is, we use a similar trick as in Lemma 15. Since we consider MPNNs in which $h(d_u) = 1$, the message function (14) indeed only depends on $\mathbf{y}$. As a consequence, M is equivalent to an anonymous MPNN. From Theorem 7 and in particular from $\mathcal{M}_{\mathit{anon}} \preceq \mathcal{M}_{\mathit{WL}}$, the proposition follows. □

The architectures $\mathcal{M}_{\mathit{dGNN}_1}$ and $\mathcal{M}_{\mathit{dGNN}_3}$ from Table 1 clearly satisfy the assumption in the previous proposition and hence $\mathcal{M}_{\mathit{dGNN}_1}, \mathcal{M}_{\mathit{dGNN}_3} \preceq \mathcal{M}_{\mathit{WL}}$.

We thus have shown the remaining part of the third item in Theorem 12. □

J. Proof of Proposition 18

The proof consists of a number of counterexamples related to the various classes of dMPNNs under consideration. For convenience, we describe the counterexamples in terms of graph neural networks rather than in their dMPNN form.

We first prove the proposition for classes of dMPNNs related to graph neural networks of the form:

$$\mathbf{L}^{(t)} := \sigma\left(\mathsf{diag}(\mathbf{g})\mathbf{A}\mathsf{diag}(\mathbf{h})\mathbf{L}^{(t-1)}\mathbf{W}^{(t)} + \mathbf{B}^{(t)}\right).$$

This includes $\mathcal{M}_{\mathit{dGNN}_i}$, for $i = 1, 2$. Consider the labelled graph $(G_1, \boldsymbol{\nu})$ with vertex labelling $\boldsymbol{\nu}_{v_1} = (1,0,0)$, $\boldsymbol{\nu}_{v_2} = \boldsymbol{\nu}_{v_3} = (0,1,0)$ and $\boldsymbol{\nu}_{v_4} = (0,0,1)$, and edges $\{v_1, v_2\}$, $\{v_1, v_3\}$, $\{v_4, v_2\}$ and $\{v_4, v_3\}$, as depicted in Figure 4.

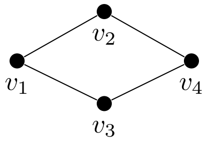

Figure 4. Graph G_1.

By definition, $\mathbf{L}^{(0)} := \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$. We note that

$$(\boldsymbol{\ell}_{M_{\mathit{WL}}}^{(1)})_{v_1} = \textsc{Hash}((1,0,0), \{\!\{(0,1,0),(0,1,0)\}\!\}) \neq (\boldsymbol{\ell}_{M_{\mathit{WL}}}^{(1)})_{v_4} = \textsc{Hash}((0,0,1), \{\!\{(0,1,0),(0,1,0)\}\!\}).$$

We next show that there exist no $\mathbf{W}^{(1)}, \mathbf{B}^{(1)}$ such that $\mathbf{L}^{(1)} \sqsubseteq \boldsymbol{\ell}^{(1)}_{M_{WL}}$. Indeed, since the degree of all vertices is 2 the computation is quite simple

$$\begin{aligned}
\mathbf{L}^{(1)} &:= \sigma\left(\mathsf{diag}(\mathbf{g})\begin{pmatrix}0 & 1 & 1 & 0\\ 1 & 0 & 0 & 1\\ 1 & 0 & 0 & 1\\ 0 & 1 & 1 & 0\end{pmatrix}\mathsf{diag}(\mathbf{h})\mathbf{L}^{(0)}\mathbf{W}^{(1)} + \mathbf{B}^{(1)}\right)\\
&= \sigma\left(\begin{pmatrix}0 & g(2)h(2) & g(2)h(2) & 0\\ g(2)h(2) & 0 & 0 & g(2)h(2)\\ g(2)h(2) & 0 & 0 & g(2)h(2)\\ 0 & g(2)h(2) & g(2)h(2) & 0\end{pmatrix}\begin{pmatrix}1 & 0 & 0\\ 0 & 1 & 0\\ 0 & 1 & 0\\ 0 & 0 & 1\end{pmatrix}\mathbf{W}^{(1)} + \mathbf{B}^{(1)}\right)\\
&= \sigma\left(\begin{pmatrix}0 & 2g(2)h(2) & 0\\ g(2)h(2) & 0 & g(2)h(2)\\ g(2)h(2) & 0 & g(2)h(2)\\ 0 & 2g(2)h(2) & 0\end{pmatrix}\mathbf{W}^{(1)} + \mathbf{B}^{(1)}\right).
\end{aligned}$$

Finally, we recall that $\mathbf{B}^{(1)}$ consists of n copies of the same row. Hence, independently of the choice of $\mathbf{W}^{(1)}$ and $\mathbf{B}^{(1)}$, vertices v_1 and v_4 will be assigned the same label, and thus $\mathbf{L}^{(1)} \not\sqsubseteq \boldsymbol{\ell}^{(1)}_{M_{WL}}$.

The second class of dMPNNs we consider are those related to graph neural networks of the form:

$$\mathbf{L}^{(t)} := \sigma\left(\mathsf{diag}(\mathbf{g})(\mathbf{A}+\mathbf{I})\mathsf{diag}(\mathbf{h})\mathbf{L}^{(t-1)}\mathbf{W}^{(t)} + \mathbf{B}^{(t)}\right).$$

This includes $\mathcal{M}_{dGNN_i}$, for $i = 3, 4$. Indeed, consider the labelled graph $(G_2, \boldsymbol{\nu})$ with one edge $\{v_1, v_2\}$, as depicted in Figure 5, and vertex labelling $\boldsymbol{\nu}_{v_1} = (1,0)$ and $\boldsymbol{\nu}_{v_2} = (0,1)$.

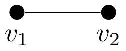

Figure 5. Graph G_2.

By definition, $\mathbf{L}^{(0)} := \left(\begin{smallmatrix}1 & 0\\ 0 & 1\end{smallmatrix}\right)$. We also note that

$$(\boldsymbol{\ell}^{(1)}_{M_{WL}})_{v_1} = \textsc{Hash}((1,0), \{\!\!\{(0,1)\}\!\!\}) \neq (\boldsymbol{\ell}^{(1)}_{M_{WL}})_{v_2} = \textsc{Hash}((0,1), \{\!\!\{(1,0)\}\!\!\}).$$

We next show that there exist no $\mathbf{W}^{(1)}, \mathbf{B}^{(1)}$ such that $\mathbf{L}^{(1)} \sqsubseteq \boldsymbol{\ell}^{(1)}_{M_{WL}}$. Indeed,

$$\begin{aligned}
\mathbf{F}^{(1)} &:= \sigma\left(\mathsf{diag}(\mathbf{g})\left(\begin{pmatrix}0 & 1\\ 1 & 0\end{pmatrix} + \begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}\right)\mathsf{diag}(\mathbf{h})\mathbf{L}^{(0)}\mathbf{W}^{(1)} + \mathbf{B}^{(1)}\right)\\
&= \sigma\left(\begin{pmatrix}g(1) & 0\\ 0 & g(1)\end{pmatrix}\begin{pmatrix}1 & 1\\ 1 & 1\end{pmatrix}\begin{pmatrix}h(1) & 0\\ 0 & h(1)\end{pmatrix}\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}\mathbf{W}^{(1)} + \mathbf{B}^{(1)}\right)\\
&= \sigma\left(\begin{pmatrix}g(1)h(1) & g(1)h(1)\\ g(1)h(1) & g(1)g(1)\end{pmatrix}\mathbf{W}^{(1)} + \mathbf{B}^{(1)}\right).
\end{aligned}$$

Hence, independently of the choice of $\mathbf{W}^{(1)}$ and $\mathbf{B}^{(1)}$, both vertices will be assigned the same label, and thus $\mathbf{L}^{(1)} \not\sqsubseteq \boldsymbol{\ell}^{(1)}_{M_{WL}}$.

Finally, we deal with the class $\mathcal{M}_{dGNN_5}$, i.e., dMPNNs related to graph neural networks of the form

$$\mathbf{L}^{(t)} := \sigma\left((\mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2} + \mathbf{I})\mathbf{L}^{(t-1)}\mathbf{W}^{(t)} + \mathbf{B}^{(t)}\right).$$

We consider the labelled graph $(G_3, \boldsymbol{\nu})$ with vertex labelling $\boldsymbol{\nu}_{v_1} = \boldsymbol{\nu}_{w_2} = \boldsymbol{\nu}_{w_3} = (1,0,0)$, $\boldsymbol{\nu}_{w_1} = \boldsymbol{\nu}_{v_2} = \boldsymbol{\nu}_{v_3} = (0,1,0)$ and $\boldsymbol{\nu}_{v_4} = \boldsymbol{\nu}_{v_5} = \boldsymbol{\nu}_{w_4} = \boldsymbol{\nu}_{w_5} = (0,0,1)$ and edges $\{v_1, v_2\}$, $\{v_1, v_3\}$, $\{v_1, v_4\}$, $\{v_1, v_5\}$ and $\{w_1, w_2\}$, $\{w_1, w_3\}$, $\{w_1, w_4\}$, $\{w_1, w_5\}$, as depicted in Figure 6.

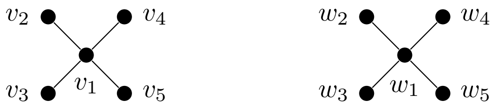

Figure 6. Graph G_3.

By definition, $\mathbf{L}^{(0)} := \begin{pmatrix} 1&0&0\\0&1&0\\0&1&0\\0&0&1\\0&0&1\\0&1&0\\1&0&0\\1&0&0\\0&0&1\\0&0&1 \end{pmatrix}$. We also note that

$$\begin{aligned}(\boldsymbol{\ell}^{(1)}_{M_{\mathit{WL}}})_{v_1} &= \textsc{Hash}((1,0,0), \{\!\{(0,1,0),(0,1,0),(0,0,1),(0,0,1)\}\!\})\\ \neq (\boldsymbol{\ell}^{(1)}_{M_{\mathit{WL}}})_{w_1} &= \textsc{Hash}((0,1,0), \{\!\{(1,0,0),(1,0,0),(0,0,1),(0,0,1)\}\!\}).\end{aligned}$$

We next show that there exist no $\mathbf{W}^{(1)}, \mathbf{B}^{(1)}$ such that $\mathbf{L}^{(1)} \sqsubseteq \boldsymbol{\ell}^{(1)}_{M_{\mathit{WL}}}$. Indeed,

$$\mathbf{L}^{(1)} := \sigma\left(\left(\mathsf{diag}\left(\begin{pmatrix}\frac{1}{2}\\1\\1\\1\\1\\\frac{1}{2}\\1\\1\\1\\1\end{pmatrix}\right)\begin{pmatrix}0&1&1&1&1&0&0&0&0&0\\1&0&0&0&0&0&0&0&0&0\\1&0&0&0&0&0&0&0&0&0\\1&0&0&0&0&0&0&0&0&0\\1&0&0&0&0&0&0&0&0&0\\0&0&0&0&0&0&1&1&1&1\\0&0&0&0&0&1&0&0&0&0\\0&0&0&0&0&1&0&0&0&0\\0&0&0&0&0&1&0&0&0&0\\0&0&0&0&0&1&0&0&0&0\end{pmatrix}\mathsf{diag}\left(\begin{pmatrix}\frac{1}{2}\\1\\1\\1\\1\\\frac{1}{2}\\1\\1\\1\\1\end{pmatrix}\right) + \mathbf{I}\right)\mathbf{L}^{(0)}\mathbf{W}^{(1)} + \mathbf{B}^{(1)}\right)$$

$$= \sigma\left(\begin{pmatrix}0&\frac{1}{2}&\frac{1}{2}&\frac{1}{2}&\frac{1}{2}&0&0&0&0&0\\\frac{1}{2}&0&0&0&0&0&0&0&0&0\\\frac{1}{2}&0&0&0&0&0&0&0&0&0\\\frac{1}{2}&0&0&0&0&0&0&0&0&0\\\frac{1}{2}&0&0&0&0&0&0&0&0&0\\0&0&0&0&0&0&\frac{1}{2}&\frac{1}{2}&\frac{1}{2}&\frac{1}{2}\\0&0&0&0&0&\frac{1}{2}&0&0&0&0\\0&0&0&0&0&\frac{1}{2}&0&0&0&0\\0&0&0&0&0&\frac{1}{2}&0&0&0&0\\0&0&0&0&0&\frac{1}{2}&0&0&0&0\end{pmatrix}\begin{pmatrix}1&0&0\\0&1&0\\0&1&0\\0&0&1\\0&0&1\\0&1&0\\1&0&0\\1&0&0\\0&0&1\\0&0&1\end{pmatrix}\mathbf{W}^{(1)}\right) = \sigma\left(\begin{pmatrix}1&1&1\\\frac{1}{2}&1&0\\\frac{1}{2}&1&0\\\frac{1}{2}&0&1\\\frac{1}{2}&0&1\\1&1&1\\1&\frac{1}{2}&0\\1&\frac{1}{2}&0\\0&\frac{1}{2}&1\\0&\frac{1}{2}&1\end{pmatrix}\mathbf{W}^{(1)} + \mathbf{B}^{(1)}\right).$$

Hence, independently of the choice of $\mathbf{W}^{(1)}$ and $\mathbf{B}^{(1)}$, vertices v_1 and w_1 will be assigned the same label, and thus $\mathbf{L}^{(1)} \not\sqsubseteq \boldsymbol{\ell}^{(1)}_{M_{\mathit{WL}}}$. □

K. Proof of Proposition 19

We recall that dMPNNs in $\mathcal{M}_{\mathit{dGNN}_6}$ correspond to graph neural network architectures of the form

$$\mathbf{L}^{(t)} := \sigma(\mathsf{diag}(\mathbf{g})(\mathbf{A} + p\mathbf{I})\mathsf{diag}(\mathbf{h})\mathbf{L}^{(t-1)}\mathbf{W}^{(t)} + \mathbf{B}^{(t)}), \tag{16}$$

where $\mathsf{diag}(\mathbf{g}) = \mathsf{diag}(\mathbf{h}) = (r\mathbf{I} + (1-r)\mathbf{D})^{-1/2}$ and σ is ReLU or sign. In fact, our proof will work for any degree-determined $\mathbf{g}$ and $\mathbf{h}$.

The argument closely follows the proof of Theorem 11. More specifically, we construct a dMPNN M corresponding to (16) such that $\boldsymbol{\ell}^{(t)}_M \sqsubseteq \boldsymbol{\ell}^{(t)}_{M_{\mathit{WL}}}$ for all $t \geq 0$. The induction hypothesis is that $\boldsymbol{\ell}^{(t)}_M \sqsubseteq \boldsymbol{\ell}^{(t)}_{M_{\mathit{WL}}}$ and $\boldsymbol{\ell}^{(t)}_M$ is row-independent modulo equality. This hypothesis is clearly satisfied, by definition, for $t = 0$.

For the inductive step we assume that $\boldsymbol{\ell}_M^{(t-1)} \sqsubseteq \boldsymbol{\ell}_{M_{\mathrm{WL}}}^{(t-1)}$ and $\boldsymbol{\ell}_M^{(t-1)}$ is row-independent modulo equality. Let us define the labelling $\boldsymbol{\kappa}_M^{(t-1)}$ such that $(\boldsymbol{\kappa}_M^{(t-1)})_v := h(d_v)(\boldsymbol{\ell}_M^{(t-1)})_v$ for all vertices v.

Lemma 28. *We have that $\boldsymbol{\kappa}_M^{(t-1)} \sqsubseteq \boldsymbol{\ell}_{M_{\mathrm{WL}}}^{(t-1)}$ and $\boldsymbol{\kappa}_M^{(t-1)}$ is row-independent modulo equality.*

Proof. Suppose that there are two vertices v and w such that

$$(\boldsymbol{\kappa}_M^{(t-1)})_v = h(d_v)(\boldsymbol{\ell}_M^{(t-1)})_v = h(d_w)(\boldsymbol{\ell}_M^{(t-1)})_w = (\boldsymbol{\kappa}_M^{(t-1)})_w.$$

This implies that $(\boldsymbol{\ell}_M^{(t-1)})_v$ is a (non-zero) scalar multiple of $(\boldsymbol{\ell}_M^{(t-1)})_w$. This is only possible when $(\boldsymbol{\ell}_M^{(t-1)})_v = (\boldsymbol{\ell}_M^{(t-1)})_w$ because $\boldsymbol{\ell}_M^{(t-1)}$ is row-independent modulo equality. In other words, $\boldsymbol{\kappa}_M^{(t-1)} \sqsubseteq \boldsymbol{\ell}_M^{(t-1)} \sqsubseteq \boldsymbol{\ell}_{M_{\mathrm{sl}WL}}^{(t-1)}$. Similarly, suppose that $\boldsymbol{\kappa}_M^{(t-1)}$ is not row-independent modulo equality then, due to the definition of $\boldsymbol{\kappa}_M^{(t-1)}$, this implies that $\boldsymbol{\ell}_M^{(t-1)}$ is also not row-independent modulo equality. □

Lemma 28 gives us sufficient conditions to repeat a key part of the argument in the proof of Theorem 11. That is, we can find a matrix $\mathbf{U}^{(t)}$ such that the labelling $\boldsymbol{\mu}^{(t)} : v \mapsto \big((\mathbf{A} + p\mathbf{I})\mathsf{diag}(\mathbf{h})\mathbf{L}^{(t-1)}\mathbf{U}^{(t)}\big)_v$ satisfies $\boldsymbol{\mu}^{(t)} \sqsubseteq \boldsymbol{\ell}_{M_{\mathrm{WL}}}^{(t)}$.

We will now prove that the labelling $\boldsymbol{\lambda}^{(t)}$ defined by $\boldsymbol{\lambda}_v^{(t)} := g(d_v)\boldsymbol{\mu}_v^{(t)}$, also satisfies $\boldsymbol{\lambda}^{(t)} \sqsubseteq \boldsymbol{\ell}_{M_{\mathrm{WL}}}^{(t)}$. We remark that $\boldsymbol{\lambda}^{(t)}$ coincides with the labelling:

$$v \mapsto (\mathsf{diag}(\mathbf{g})(\mathbf{A} + p\mathbf{I})\mathsf{diag}(\mathbf{h})\mathbf{L}^{(t-1)}\mathbf{U}^{(t)})_v.$$

Lemma 29. *The exists a constant m_p, only dependent on $\mathbf{g}$ and the number n of vertices, such that $\boldsymbol{\lambda}^{(t)} \sqsubseteq \boldsymbol{\ell}_{M_{\mathrm{WL}}}^{(t)}$, for every $m_p < p < 1$.*

Proof. We will choose m_p at the end of the proof. For now suppose that $\boldsymbol{\lambda}^{(t)} \not\sqsubseteq \boldsymbol{\ell}_{M_{\mathrm{WL}}}^{(t)}$. Then there exist two vertices v and w such that

$$\boldsymbol{\lambda}_v^{(t)} = \boldsymbol{\lambda}_w^{(t)} \text{ and } (\boldsymbol{\ell}_{M_{\mathrm{WL}}}^{(t)})_v \neq (\boldsymbol{\ell}_{M_{\mathrm{WL}}}^{(t)})_w.$$

The latter implies that $\boldsymbol{\mu}_v^{(t)} \neq \boldsymbol{\mu}_w^{(t)}$ and thus $\boldsymbol{\lambda}_v^{(t)} = \boldsymbol{\lambda}_w^{(t)}$ implies that $g(d_v) \neq g(d_w)$.

We recall some facts from the proof of Theorem 11, and from equations (10) and (11) in particular. An entry in $\boldsymbol{\mu}_v^{(t)}$ is either 0 or $1, 2, \ldots, n$ or $i + p$, for some $i \in \{0, 1, \ldots, n\}$. Furthermore, at least one entry must be distinct from 0. Also, $\boldsymbol{\lambda}_v^{(t)} = \boldsymbol{\lambda}_w^{(t)}$ implies that the positions of the non-zero entries in $\boldsymbol{\mu}_v^{(t)}$ and $\boldsymbol{\mu}_w^{(t)}$ coincide. (Recall that the image of g is $\mathbb{A}^+$). Let Z be the positions in $\boldsymbol{\mu}_v^{(t)}$ (and thus also in $\boldsymbol{\mu}_w^{(t)}$) that carry non-zero values.

We can now infer that $\boldsymbol{\lambda}_v^{(t)} = \boldsymbol{\lambda}_w^{(t)}$ implies that for every $i \in Z$:

$$\frac{\boldsymbol{\mu}_{vi}^{(t)}}{\boldsymbol{\mu}_{wi}^{(t)}} = \frac{g(d_w)}{g(d_v)} \neq 1.$$

Moreover, both in $\boldsymbol{\mu}_v^{(t)}$ and $\boldsymbol{\mu}_w^{(t)}$ there are unique positions i_1 and i_2, respectively, whose corresponding entry contain p. We now consider three cases:

$$\text{(a) } \frac{\boldsymbol{\mu}_{vi_1}^{(t)}}{\boldsymbol{\mu}_{wi_1}^{(t)}} = \frac{i+p}{j}; \quad \text{(b) } \frac{\boldsymbol{\mu}_{vi_2}^{(t)}}{\boldsymbol{\mu}_{wi_2}^{(t)}} = \frac{i}{j+p}; \quad \text{(c) } \frac{\boldsymbol{\mu}_{vi_1}^{(t)}}{\boldsymbol{\mu}_{wi_1}^{(t)}} = \frac{i+p}{j+p} \text{ (this is the case if and only if } i_1 = i_2\text{)},$$

for some $i, j \in \{0, 1, 2, \ldots, n\}$. To define m_p, let $\Gamma := \left\{ \frac{g(d_w)}{g(d_v)} \;\middle|\; g(d_v) \neq g(d_w) \text{ and } v, w \in V \right\}$ and consider

$$P_a := \left\{ \alpha j - i \;\middle|\; 0 \leq \alpha j - i < 1, i, j \in \{0, 1, 2 \ldots, n\}, \alpha \in \Gamma \right\}$$
$$P_b := \left\{ \frac{i - \alpha j}{\alpha} \;\middle|\; 0 \leq \frac{i - \alpha j}{\alpha} < 1, i, j \in \{0, 1, 2, \ldots, n\}, \alpha \in \Gamma \right\}$$

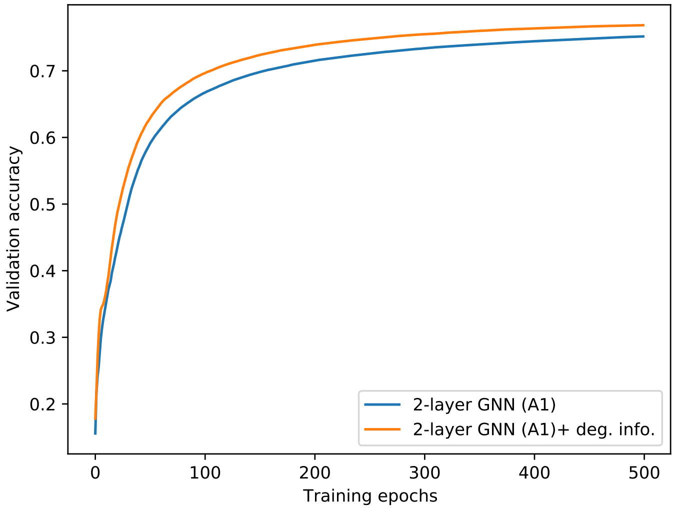

Figure 7. Accuracy of two 2-layer GNN with Arch. (A1) on the validation set plotted for the Citeseer dataset (and a copy enriched with degree information) for increasing training epochs

$$P_c := \left\{ \frac{\alpha j - i}{1-\alpha} \,\middle|\, 0 \leq \frac{\alpha(j-i)}{1-\alpha} < 1, i, j \in \{0,1,2,\dots,n\}, \alpha \in \Gamma \right\}.$$

We define $m_p = \max\{P_1 \cup P_2 \cup P_3 \cup \{0\}\}$ and we claim that for all p satisfying $m_p < p < 1$ the lemma holds.

By definition of P_a, $\alpha j - i \neq p$ and thus $\frac{i+p}{j} \neq \alpha$ for any $\alpha \in \Gamma$ and $i, j \in \{0,1,\dots,n\}$. This rules out (a). Similarly, by definition of P_b, $\frac{i-\alpha j}{\alpha} \neq p$ and thus $\frac{i}{j+p} \neq \alpha$ for any $\alpha \in \Gamma$ $i, j \in \{0,1,\dots,n\}$. This rules out (b). Finally, by definition of P_3, $\frac{\alpha j - i}{1-\alpha} \neq p$ and thus $\frac{i+p}{j+p} \neq \alpha$ for any $\alpha \in \Gamma$ $i, j \in \{0,1,\dots,n\}$. This rules out (c). We conclude, as our initial assumption cannot be valid for this m_p. □

From here, we can again follow the proof of Theorem 11 to construct a matrix $\mathbf{X}^{(t)}$ such that the labelling $\boldsymbol{\ell}_M^{(t)}$ defined by $\sigma(\mathsf{diag}(\mathbf{g})(\mathbf{A} + p\mathbf{I})\mathsf{diag}(\mathbf{h})\mathbf{L}^{(t-1)}\mathbf{U}^{(t)}\mathbf{X}^{(t)} + \mathbf{B}^{(t)})$ with $\mathbf{B}^{(t)} = -\mathbf{J}$ if σ is the sign function, and $\mathbf{B}^{(t)} = -q\mathbf{J}$ if σ is the ReLU function, is such that $\boldsymbol{\ell}_M^{(t)} \sqsubseteq \boldsymbol{\ell}_{M_{\mathsf{WL}}}^{(t)}$ and $\boldsymbol{\ell}_M^{(t)}$ is row-independent modulo equality. This concludes the proof for dMPNNs arising from graph neural networks of the form (16). □

L. Experiments

We explore some practical repercussions of our theoretical analysis. Since the degree-aware MPPNs listed in Table 1 have been experimentally validated already in the literature, we here focus on anonymous MPNNs. In particular, we address the following two questions:

Q1 Are the more succinct ReLU-GNNs (4) competitive to the ones proposed in Morris et al. (2019)? Recall that our ReLU-GNNs require considerably less parameters and need half the number of layers to simulate WL (Theorem 11).

Q2 Do "anonymous" GNNs benefit from the inclusion of degree information in the initial vertex features? Recall that Proposition 14 implies that the power of degree-aware MPNNs can be matched with that of anonymous MPNNs, provided that degree information is added to the initial features. Furthermore, Proposition 16 indicates that such degree information is best added before aggregation, e.g., in the initial features.

Baselines and datasets. We have implemented several GNN variants in Tensorflow 2.0 (Abadi et al., 2016), including the WL-poweful ReLU-GNNs from Morris et al. (2019). Our implementation is available here: `https://github.com/gaperez64/gnns`. For the datasets, we considered Citeseer, Cora, and Pubmed citation networks (Sen et al., 2008) as pre-processed in the implementation of Kipf & Welling (2017). During training, all feature vectors and 20 labels per vertex

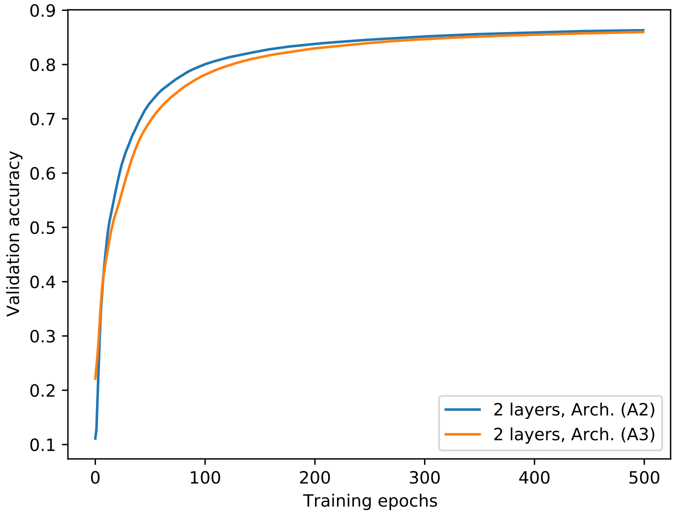

Figure 8. Accuracy of a 2-layer GNN with Arch. (A2) and a 2-layer GNN with Arch. (A3) on the validation set plotted for the Cora dataset for increasing training epochs

Architecture	**Citeseer**	**Cora**	**Pubmed**
Arch. (A1)	75.144	84.881	56.884
Arch. (A1)*	76.829	73.650	72.275
Arch. (A2)	81.046	86.310	87.923
Arch. (A3)	79.562	85.980	87.321

Table 2. Classification accuracy (percentage) on the test subsets of the considered datasets; * indicates that the model was trained and evaluated against datasets with degree information

class were used. Finally we added copies of the datasets with feature vectors extended with a single value: the degree of the vertex.

Experimental setup. We report on the prediction accuracy of 2-layer GNNs with the following architectures, where Architecture (A1) corresponds to "vanilla" GNNs; (A2) is the double-weight-matrix GNN architecture used in Morris et al. (2019) to prove Theorem 10; (A3) is our newly proposed architecture from the proof of Theorem 11.

$$\mathrm{ReLU}(\mathbf{A}\mathbf{L}^{(t-1)}\mathbf{W}^{(t)} + \mathbf{B}^{(t)}) \tag{A1}$$
$$\mathrm{ReLU}(\mathbf{A}\mathbf{L}^{(t-1)}\mathbf{W}_1^{(t)} + \mathbf{L}^{(t-1)}\mathbf{W}_2^{(t)} + \mathbf{B}^{(t)}) \tag{A2}$$
$$\mathrm{ReLU}((\mathbf{A} + p\mathbf{I})\mathbf{L}^{(t-1)}\mathbf{W}^{(t)} + \mathbf{B}^{(t)}) \tag{A3}$$

They are all trained using 1K labelled examples. As in Kipf & Welling (2017), we use the data splits given in Yang et al. (2016) with an extra validation set with 500 labelled examples. We use a dropout rate of 0.5; L2-regularization factor of $5 \cdot 10^{-4}$; and 16 hidden units, for all layers. All models are trained for 500 epochs using Adam (Kingma & Ba, 2015) with a learning rate of 0.01. The matrix weights are initialized following Glorot & Bengio (2010).

Results and discussion. We first observe that, as can be seen in Figure 7 and Table 2, degree information does increase the accuracy of (same architecture) GNNs in $2/3$ datasets. Secondly, Figure 8 and Table 2 confirm that ReLU-GNNs with less matrices have competitive accuracy if we use the scaling factor p suggested by the proof of Theorem 11. In our experiments we obtained best results by setting $p := 0.5872$.

L.1. Additional figures

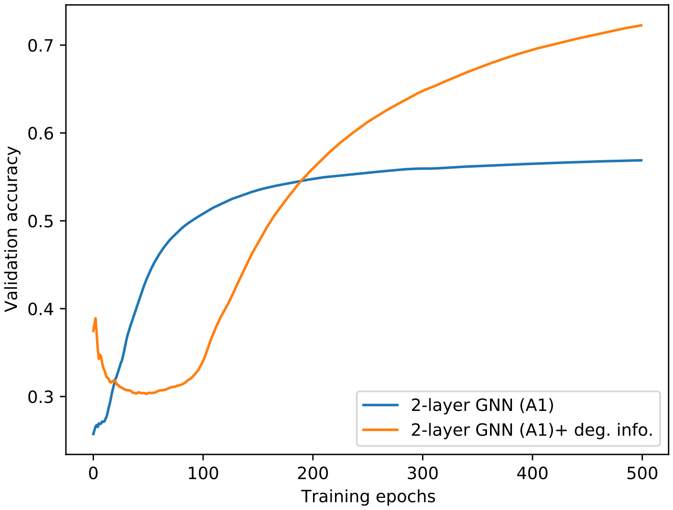

Figure 9. Same information as in Figure 7 but for the Pubmed dataset and a degree-information enriched copy

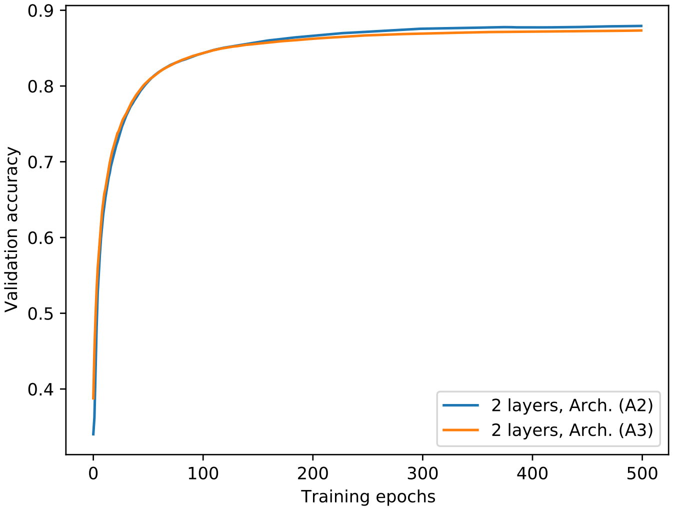

Figure 10. Same information as in Figure 8 but for the Pubmed dataset