

---

# Distance Encoding: Design Provably More Powerful Neural Networks for Graph Representation Learning

---

**Pan Li**

Department of Computer Science  
Purdue University  
panli@purdue.edu

**Yanbang Wang**

Department of Computer Science  
Stanford University  
ywangdr@cs.stanford.edu

**Hongwei Wang**

Department of Computer Science  
Stanford University  
hongweiw@cs.stanford.edu

**Jure Leskovec**

Department of Computer Science  
Stanford University  
jure@cs.stanford.edu

## Abstract

Learning representations of sets of nodes in a graph is crucial for applications ranging from node-role discovery to link prediction and molecule classification. Graph Neural Networks (GNNs) have achieved great success in graph representation learning. However, expressive power of GNNs is limited by the 1-Weisfeiler-Lehman (WL) test and thus GNNs generate identical representations for graph substructures that may in fact be very different. More powerful GNNs, proposed recently by mimicking higher-order-WL tests, only focus on representing entire graphs and they are computationally inefficient as they cannot utilize sparsity of the underlying graph. Here we propose and mathematically analyze a general class of structure-related features, termed Distance Encoding (DE). DE assists GNNs in representing any set of nodes, while providing strictly more expressive power than the 1-WL test. DE captures the distance between the node set whose representation is to be learned and each node in the graph. To capture the distance DE can apply various graph-distance measures such as shortest path distance or generalized PageRank scores. We propose two ways for GNNs to use DEs (1) as extra node features, and (2) as controllers of message aggregation in GNNs. Both approaches can utilize the sparse structure of the underlying graph, which leads to computational efficiency and scalability. We also prove that DE can distinguish node sets embedded in almost all regular graphs where traditional GNNs always fail. We evaluate DE on three tasks over six real networks: structural role prediction, link prediction, and triangle prediction. Results show that our models outperform GNNs without DE by up-to 15% in accuracy and AUROC. Furthermore, our models also significantly outperform other state-of-the-art methods especially designed for the above tasks.

## 1 Introduction

Graph representation learning aims to learn representation vectors of graph-structured data [1]. Representations of node sets in a graph can be leveraged for a wide range of applications, such as discovery of functions/roles of nodes based on individual node representations [2–6], link or link type prediction based on node-pair representations [7–10] and graph comparison or molecule classification based on entire-graph representations [11–17].

Graph neural networks (GNNs), inheriting the power of neural networks [18], have become the *de facto* standard for representation learning in graphs [19]. Generally, GNNs use message passing procedure over the input graph, which can be summarized in three steps: (1) Initialize node representations with their initial attributes (if given) or structural features such as node degrees;

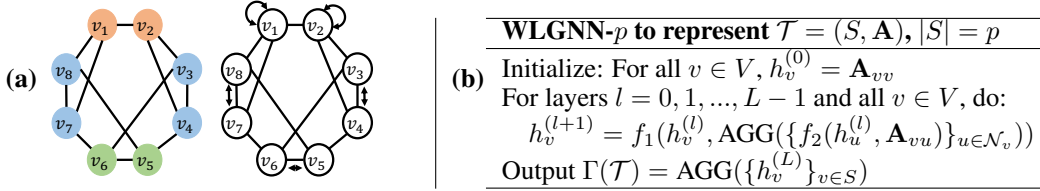


Figure 1: (a) 3-regular graph with 8 nodes. Briefly assume that all node attributes are the same and the nodes can only be distinguished based on their network structure. Then for all nodes, WLGN will produce the same representation and thus fail to distinguish them. However, nodes with different colors should have different representations, as they are not structurally equivalent (or “isomorphic” as defined in Section 2). Furthermore, WLGN cannot distinguish all the node-pairs (e.g.,  $\{v_1, v_2\}$  vs  $\{v_4, v_7\}$ ). However, if we use shortest-path-distances (SPDs) between nodes as features we can distinguish blue nodes from green and red nodes because there is another node with SPD= 3 to a blue node of interest (e.g., SPD between  $v_3$  and  $v_8$ ), while all SPDs between other nodes to red/green nodes are less than 3. Note that the structural equivalence between any two nodes of the same color can be obtained from the reflexivity of the graph while the equivalence between two vertically-aligned blue nodes can be further obtained from the node permutation shown in the right. (b) WLGN algorithm to represent a node set  $S$  of size  $p$  —  $f_i(\cdot)$ ’s are arbitrary neural networks;  $\text{AGG}(\cdot)$ ’s are set-pooling operators;  $L$  is the number of layers.

(2) Iteratively update the representation of each node by aggregating over the representations of its neighboring nodes; (3) Readout the final representation of a single node, a set of nodes, or the entire node set as required by the task. Under the above framework, researchers have proposed many GNN architectures [14–16, 20–23]. Interested readers may refer to tutorials on GNNs for further details [1, 24].

Despite the success of GNNs, their representation power in representation learning is limited [16]. Recent works proved that the representation power of GNNs that follow the above framework is bounded by the 1-WL test [16, 25, 26] (We shall refer to these GNNs as WLGNs). Concretely, WLGNs yield identical vector representations for any subgraph structure that the 1-WL test cannot distinguish. Consider an extreme case: If node attributes are all nodes are the same, then for any node in a  $r$ -regular graph GNN will output identical representation. Such an issue becomes even worse when WLGNs are used to extract representations of node sets, e.g., node-pairs for link prediction (Fig. 1 (a)). A few works have been recently proposed to improve the power of WLGNs [27]. However, they either focus on building theory only for entire-graph representations [26–30], or show empirical success using heuristic methods without strong theoretical characterization [9, 10, 17, 31–33]. We review these methods in detail in Section 4.

Here we address the limitations of WLGNs and propose and mathematically analyze a new class of node features, termed *Distance Encoding (DE)*. DE comes with both theoretical guarantees and empirical efficiency. Given a node set  $S$  whose structural representation is to be learnt, for every node  $u$  in the graph DE is defined as a mapping of a set of landing probabilities of random walks from each node of the set  $S$  to node  $u$ . DE may use measures such as shortest path distance (SPD) and generalized PageRank scores [34]. DE can be combined with any GNN architecture in simple but effective ways: First, we propose DE-GNN that utilizes DE as an extra node feature. We further enhance DE-GNN by allowing DE to control the message aggregation procedure of WLGNs, which yields another model DEA-GNN. Since DE purely depends on the graph structure and is independent of node identifiers, DE also provides inductive and generalization capability.

We mathematically analyze the expressive power of DE-GNN and DEA-GNN for structural representation learning. We prove that the two models are able to distinguish two non-isomorphic equally-sized node sets (including nodes, node-pairs, . . . , entire-graphs) that are embedded in almost all sparse regular graphs, where WLGN always fails to distinguish them unless discriminatory node/edge attributes are available. We also prove that the two models are not more powerful than WLGN when applied to distance regular graphs [35], which implies the limitation of DEs. However, we show that DE has an extra power to learn the structural representations of node-pairs over distance regular graphs [35].

We experimentally evaluate DE-GNN and DEA-GNN on three levels of tasks including node-structural-role classification (node-level), link prediction (node-pair-level), triangle prediction (node-triad-level). Our methods outperform WLGN on all three tasks by up-to 15% improvement in average accuracy. Our methods also outperform other baselines specifically designed for these tasks.

## 2 Preliminaries

In this section we formally define the notion of structural representation and review how WLGN learns structural representation and its relation to the 1-WL test.

### 2.1 Graph Representation Learning

**Definition 2.1.** We consider an undirected graph which can be represented as  $G = (V, E, \mathbf{A})$ , where  $V = [n]$  is the node set,  $E \subseteq V \times V$  is the edge set, and  $\mathbf{A}$  contains all features in the space  $\mathcal{A} \subset \mathbb{R}^{n \times n \times k}$ . Its diagonal component,  $\mathbf{A}_{vv}$ , denotes the node attributes of node  $v (\in V)$ , while its off-diagonal component in  $\mathbf{A}_{vu}$  denotes the node-pair attributes of  $(v, u)$ . We set  $\mathbf{A}_{vu}$  as all zeros if  $(v, u) \notin E$ . In practice, graphs are usually sparse, *i.e.*,  $|E| \ll n^2$ . We introduce  $A \in \{0, 1\}^{n \times n}$  to denote the adjacency matrix of  $G$  such that  $A_{uv} = 1$  iff  $(u, v) \in E$ . Note that  $A$  can be also viewed as one slice of the feature tensor  $\mathbf{A}$ . If no node/edge attributes are available, we let  $\mathbf{A} = A$ .

**Definition 2.2.** The node permutation denoted by  $\pi$  is a bijective mapping from  $V$  to  $V$ . All possible  $\pi$ 's are collected in the permutation group  $\Pi_n$ . We denote  $\pi$  acting on a subset  $S$  of  $V$  as  $\pi(S) = \{\pi(i) | i \in S\}$ . We further define  $\pi(\mathbf{A})_{uv} = \mathbf{A}_{\pi^{-1}(u)\pi^{-1}(v)}$  for any  $(u, v) \in V \times V$ .

**Definition 2.3.** Denote all  $p$ -sized subsets  $S$  of  $V$  as  $S \in \mathcal{P}_p(V)$  and define the space  $\Omega_p = \mathcal{P}_p(V) \times \mathcal{A}$ . For two tuples  $\mathcal{T}_1 = (S^{(1)}, \mathbf{A}^{(1)})$  and  $\mathcal{T}_2 = (S^{(2)}, \mathbf{A}^{(2)})$  in  $\Omega_p$ , we call that they are *isomorphic* (otherwise *non-isomorphic*), if  $\exists \pi \in \Pi_n$  such that  $S^{(1)} = \pi(S^{(2)})$  and  $\mathbf{A}^{(1)} = \pi(\mathbf{A}^{(2)})$ .

**Definition 2.4.** A function  $f$  defined on  $\Omega_p$  is *invariant* if  $\forall \pi \in \Pi_n, f(S, \mathbf{A}) = f(\pi(S), \pi(\mathbf{A}))$ .

**Definition 2.5.** The *structural representation* of a tuple  $(S, \mathbf{A})$  is an invariant function  $\Gamma(\cdot) : \Omega_p \rightarrow \mathbb{R}^d$  where  $d$  is the dimension of representation. Therefore, if two tuples are isomorphic, they should have the same structural representation.

The invariant property is critical for the inductive and generalization capability as it frees structural representations from node identifiers and effectively reduces the problem dimension by incorporating the symmetry of the parameter space [29] (*e.g.*, the convolutional layers in GCN [20]). The invariant property also implies that structural representations do not allow encoding the absolute positions of  $S$  in the graph.

The definition of structural representation is very general. Suppose we set two node sets  $S^{(1)}, S^{(2)}$  as two single nodes and set two graph structures  $\mathbf{A}^{(1)}$  and  $\mathbf{A}^{(2)}$  as the ego-networks around these two nodes. Then, the definition of structural representation provides a mathematical characterization the concept ‘‘structural roles’’ of nodes [3, 5, 6], where two far-away nodes could have the same structural roles (representations) as long as their ego-networks have the same structure.

Note that depending on the application one can vary/select the size  $p$  of the node set  $S$ . For example, when  $p = 1$  then we are in the regime of node classification,  $p = 2$  is link prediction, and when  $S = V$ , structural representations reduce to entire graph representations. However, in this work we will primarily focus on the case that the node set  $S$  has a fixed and small size  $p$ , where  $p$  does not depend on the graph size  $n$ . Although Corollary 3.4 later shows the potential of our techniques on learning the entire graph representations, this is not the main focus of our work here. We expect the techniques proposed here can be further used for entire-graph representations while we leave the detailed investigation for future work.

Although structural representation defines a more general concept, it shares some properties with traditional entire-graph representation. For example, the universal approximation theorem regarding entire-graph representation [30] can be directly generalized to the case of structural representations:

**Theorem 2.6.** If structural representations  $\Gamma$  are different over any two non-isomorphic tuples  $\mathcal{T}_1$  and  $\mathcal{T}_2$  in  $\Omega_p$ , then for any invariant function  $f : \Omega_p \rightarrow \mathbb{R}$ ,  $f$  can be universally approximated by feeding  $\Gamma$  into a 3-layer feed-forward neural network with ReLU as the activation function, as long as (1) the feature space  $\mathcal{A}$  is compact and (2)  $f(S, \cdot)$  is continuous over  $\mathcal{A}$  for any  $S \in \mathcal{P}_p(V)$ .

Theorem 2.6 formally establishes the relation between learning structural representations and distinguishing non-isomorphic structures, *i.e.*,  $\Gamma(\mathcal{T}_1) \neq \Gamma(\mathcal{T}_2)$  iff  $\mathcal{T}_1$  and  $\mathcal{T}_2$  are non-isomorphic. However, no polynomial algorithm has been found to distinguish even just non-isomorphic entire graphs ( $S = V$ ) without node/edge attributes ( $\mathbf{A} = A$ ), which is known as the graph isomorphism problem [36]. In this work, we will use the range of non-isomorphic structures that GNNs can distinguish to characterize their expressive power for graph representation learning.

## 2.2 Weisfeiler-Lehman Tests and WLGN for Structural Representation Learning

Weisfeiler-Lehman test (WL-test) is a family of very successful algorithmic heuristics used in graph isomorphism problems [25]. 1-WL test, the simplest one among this family, starts with coloring nodes with their degrees, then it iteratively aggregates the colors of nodes and their neighborhoods, and hashes the aggregated colors into unique new colors. The coloring procedure finally converges to some static node-color configuration. Here a node-color configuration is a multiset that records the types of colors and their numbers. Different node-color configurations indicate two graphs are non-isomorphic while the reverse statement is not always true.

More than the graph isomorphism problem, the node colors obtained by the 1-WL test naturally provide a test of structural isomorphism. Consider two tuples  $\mathcal{T}_1 = (S^{(1)}, \mathbf{A}^{(1)})$  and  $\mathcal{T}_2 = (S^{(2)}, \mathbf{A}^{(2)})$  according to Definition 2.3. We temporarily ignore node/edge attributes for simplicity, so  $\mathbf{A}^{(1)}, \mathbf{A}^{(2)}$  reduce to adjacent matrices. It is easy to show that different node-color configurations of nodes in  $S^{(1)}$  and in  $S^{(2)}$  obtained by the 1-WL test also indicate that  $\mathcal{T}_1$  and  $\mathcal{T}_2$  are not isomorphic.

WLGNNs refer to those GNNs that mimic the 1-WL test to learn structural representation, which is summarized in Fig. 1 (b). It covers many well-known GNNs of which difference may appear in the implementation of neural networks  $f_i$  and set-poolings  $\text{AGG}(\cdot)$  (Fig. 1 (b)), including GCN [20], GraphSAGE [21], GAT [22], MPNN [14], GIN [16] and many others [29]. Note that we use  $\text{WLGNN-}p$  to denote the WLGNN that is to learn structural representations of node sets  $S$  with size  $|S| = p$ . One may directly choose  $S = V$  to obtain the entire-graph representation. Theoretically, the structural representation power of  $\text{WLGNN-}p$  is provably bounded by the 1-WL test [16]. The result can be also generalized to the case of structural representations as follows.

**Theorem 2.7.** Consider two tuples  $\mathcal{T}_1 = (S^{(1)}, \mathbf{A}^{(1)})$  and  $\mathcal{T}_2 = (S^{(2)}, \mathbf{A}^{(2)})$  in  $\Omega_p$ . If  $\mathcal{T}_1, \mathcal{T}_2$  cannot be distinguished by the 1-WL test, then the corresponding outputs of  $\text{WLGNN-}p$  satisfy  $\Gamma(\mathcal{T}_1) = \Gamma(\mathcal{T}_2)$ . On the other side, if they can be distinguished by the 1-WL test and we suppose aggregation operations (AGG) and neural networks  $f_1, f_2$  are all injective mappings, then with a large enough number of layers  $L$ , the outputs of  $\text{WLGNN-}p$  also satisfy  $\Gamma(\mathcal{T}_1) \neq \Gamma(\mathcal{T}_2)$ .

Because of Theorem 2.7, WLGNN inherits the limitation of the 1-WL test. For example, WLGNN cannot distinguish two equal-sized node sets in all  $r$ -regular graphs (unless node/edge features are discriminatory). Here, a  $r$ -regular graph means that all its nodes have degree  $r$ . Therefore, researchers have recently focused on designing GNNs with expressive power greater than the 1-WL test. Here we will improve the power of GNNs by developing a general class of structural features.

## 3 Distance Encoding and Its Power

### 3.1 Distance Encoding

Suppose we aim to *learn the structural representation of the target node set  $S$* . Intuitively, our proposed DE will then encode the distance from  $S$  to any other node  $u$ . We define DE as follows:

**Definition 3.1.** Given a target set of nodes  $S \in 2^V \setminus \emptyset$  of  $G$  with the adjacency matrix  $A$ , we denote *distance encoding* as a function  $\zeta(\cdot|S, A) : V \rightarrow \mathbb{R}^k$ .  $\zeta$  should also be permutation invariant, *i.e.*,  $\zeta(u|S, A) = \zeta(\pi(u)|\pi(S), \pi(A))$  for all  $u \in V$  and  $\pi \in \Pi_n$ . Then we denote DEs *w.r.t.* the size of  $S$  and call them as *DE- $p$*  if  $|S| = p$ .

Later we use  $\zeta(u|S)$  for brevity where  $A$  could be inferred from the context. For simplicity, we choose DE as a set aggregation (*e.g.*, the sum-pooling) of DEs between nodes  $u, v$  where  $v \in S$ :

$$\zeta(u|S) = \text{AGG}(\{\zeta(u|v)|v \in S\}) \quad (1)$$

More complicated DE may be used while this simple design can be efficiently implemented and achieves good empirical performance. Then, the problem reduces to choosing a proper  $\zeta(u|v)$ . Again for simplicity, we consider the following class of functions that is based on the mapping of a list of landing probabilities of random walks from  $v$  to  $u$  over the graph, *i.e.*,

$$\zeta(u|v) = f_3(\ell_{uv}), \ell_{uv} = ((W)_{uv}, (W^2)_{uv}, \dots, (W^k)_{uv}, \dots) \quad (2)$$

where  $W = AD^{-1}$  is the random walk matrix,  $f_3$  may be simply designed by some heuristics or be parameterized and learnt as a feed-forward neural network. In practice, a finite length of  $\ell_{vu}$ , say 3,4, is enough. Note that Eq. (2) covers many important distance measures. First, setting  $f_3(\ell_{uv})$  as

the first non-zero position in  $\ell_{uv}$  gives the *shortest-path-distance (SPD)* from  $v$  to  $u$ . We denote this specific choice as  $\zeta_{spd}(u|v)$ . Second, one may also use *generalized PageRank scores* [34]:

$$\zeta_{gpr}(u|v) = \sum_{k \geq 1} \gamma_k (W^k)_{uv} = \left( \sum_{k \geq 1} \gamma_k W^k \right)_{uv}, \quad \gamma_k \in \mathbb{R}, \text{ for all } k \in \mathbb{N}. \quad (3)$$

Note that the permutation invariant property of DE is beneficial for inductive learning, which fundamentally differs from positional node embeddings such as node2vec [37] or one-hot node identifiers. In the rest of this work, we will show that DE improves the expressive power of GNNs in both theory and practice. In Section 3.2, we use DE as extra node features. We term this model as DE-GNN, and theoretically demonstrate its expressive power. In the next subsection, we further use DE-1 to control the aggregation procedure of WLGNN. We term this model as DEA-GNN and extend our theory there.

### 3.2 DE-GNN— Distance Encodings as Node Features

DE can be used as extra node features. Specifically, we improve WLGNNs by setting  $h_v^{(0)} = \mathbf{A}_{vv} \oplus \zeta(v|S)$  where  $\oplus$  is the concatenation. We call the obtained model DE-GNN. We similarly use **DE-GNN- $p$**  to specify the case when  $|S| = p$ . For simplicity, we give the following definition.

**Definition 3.2.** DE-GNN is called *proper* if  $f_1, f_2$ , AGGs in the WLGNN (Fig. 1 (b)), and AGG in Eq. (1),  $f_3$  in Eq. (2) are injective mappings as long as the input features are all countable.

We know that a proper DE-GNN exists because of the universal approximation theorem of feed-forward networks (to construct  $f_i, i \in \{1, 2, 3\}$ ) and Deep Sets [38] (to construct AGGs).

#### 3.2.1 The Expressive Power of DE-GNN

Next, we demonstrate the power of DE-GNN to distinguish structural representations. Recall that the fundamental limit of WLGNN is the 1-WL test for structural representation (Theorem 2.7). One important class of graphs that cannot be distinguished by the 1-WL test are regular graphs (although, in practice, node/edge attributes may help diminish such difficulty by breaking the symmetry). In theory, we may consider the most difficult case by assuming that no node/edge attributes are available. In the following, our main theorem shows that even in the most difficult case, DE-GNN is able to distinguish two equal-sized node sets that are embedded in almost all  $r$ -regular graphs. One example where DE-GNN using  $\zeta_{spd}(\cdot)$  (SPD) is shown in Fig. 1 (a): The blue nodes can be easily distinguished from the green or red nodes as SPD= 3 may appear between two nodes when a blue node is the node set of interest, while all SPDs from other nodes to red and green nodes are less than 3. Actually, DE-GNN-1 with SPD may also distinguish the red or green nodes by investigating its procedure in details (Fig. 3 in Appendix).

**Theorem 3.3.** Given two equal-sized sets  $S^{(1)}, S^{(2)} \subset V$ ,  $|S^{(1)}| = |S^{(2)}| = p$ . Consider two tuples  $\mathcal{T}^{(1)} = (S^{(1)}, \mathbf{A}^{(1)})$  and  $\mathcal{T}^{(2)} = (S^{(2)}, \mathbf{A}^{(2)})$  in the most difficult setting where features  $\mathbf{A}^{(1)}$  and  $\mathbf{A}^{(2)}$  are only different in graph structures specified by  $A^{(1)}$  and  $A^{(2)}$  respectively. Suppose  $A^{(1)}$  and  $A^{(2)}$  are uniformly independently sampled from all  $r$ -regular graphs over  $V$  where  $3 \leq r < (2 \log n)^{1/2}$ . Then, for any small constant  $\epsilon > 0$ , within  $L \leq \lceil (\frac{1}{2} + \epsilon) \frac{\log n}{\log(r-1)} \rceil$  layers, there exists a proper DE-GNN- $p$  using DEs  $\zeta(u|S^{(1)})$ ,  $\zeta(u|S^{(2)})$  for all  $u \in V$ , such that with probability  $1 - o(n^{-1})$ , the outputs  $\Gamma(\mathcal{T}^{(1)}) \neq \Gamma(\mathcal{T}^{(2)})$ . Specifically,  $f_3$  can be simply chosen as SPD, *i.e.*,  $\zeta(u|v) = \zeta_{spd}(u|v)$ . The big-O notations here and later are w.r.t.  $n$ .

*Remark 3.1.* In some cases, we are to learn representations of structures that lie in a single large graph, *i.e.*,  $A^{(1)} = A^{(2)}$ . Actually, there is no additional difficulty to extend the proof of Theorem 3.3 to this setting as long as  $A^{(1)} (= A^{(2)})$  is uniformly sampled from all  $r$ -regular graphs and  $S^{(1)} \cap S^{(2)} = \emptyset$ . The underlying intuition is that for large  $n$ , the local subgraphs (within  $L$ -hop neighbors) around two non-overlapping fixed sets  $S^{(1)}, S^{(2)}$  are almost independent. Simulation results to validate the single node case ( $p = 1$ ) of Theorem 3.3 and Remark 3.1 are shown in Fig. 2 (a).

Actually, the power of structural representations of small node sets can be used to further characterize the power of entire graph representations. Consider that we directly aggregate all the representations of nodes of a graph output by DE-GNN-1 via set-pooling as the graph representation, which is a common strategy adopted to learn graph representation via WLGNN-1 [14, 16, 23]. So how about the power of DE-GNN-1 to represent graphs? To answer this question, suppose two  $n$ -sized  $r$ -regular graphs  $\mathbf{A}^{(1)}$

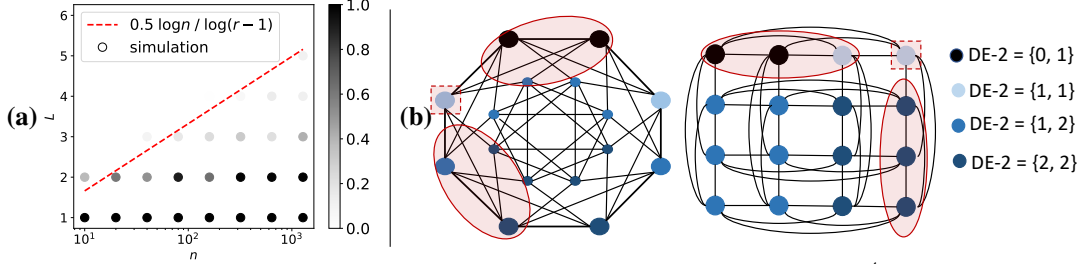


Figure 2: **(a)** Simulation to validate Theorem 3.3. We uniformly at random generate  $10^4/n$  3-regular graphs and compare node representations output by a randomly initialized but untrained DE-GNN-1 with  $L$  layers,  $L \leq 6$ . All the nodes in these graphs are considered and thus for each  $n$ , there are  $10^4$  nodes from the same or different graphs. For any two nodes  $u, v$ , if  $\|h_u^{(L)} - h_v^{(L)}\|_2$  is greater than machine accuracy, they are regarded to be distinguishable. The colors of the scatter plot indicate the portion of two nodes that are not distinguishable by DE-GNN-1. The red line is boundary predicted by our theory, which well matches the simulation. **(b)** The power of DE-2. The left is the Shrikhande graph while the right is the  $4 \times 4$  Rook's graph. DE-GNN-1 assigns all nodes with the same representation. DE-GNN-2 may distinguish the structures by learning representations of node-pairs (edges)—the node-pairs colored by black. Each node is colored with its DE-2 that is a set of SPDs to either node in the target edges (Eq. (1)). Note the neighbors of nodes with DE-2 =  $\{1, 1\}$  (covered by dashed boxes) that are highlighted by red ellipses. As these neighbors have different DE-2's, after one layer of DE-GNN-2, the intermediate representations of nodes with DE-2 =  $\{1, 1\}$  are different between these two graphs. Using another layer, DE-GNN-2 can distinguish the representations of two target edges.

and  $\mathbf{A}^{(2)}$  satisfy the condition in Theorem 3.3. Then, by using a union bound, Theorem 3.3 indicates that for a node  $v \in V$ , its representation  $\Gamma((v, \mathbf{A}^{(1)})) \notin \{\Gamma((u, \mathbf{A}^{(2)})) | u \in V\}$  with probability  $1 - no(n^{-1}) = 1 - o(1)$ . Therefore, these two graphs  $\mathbf{A}^{(1)}$  and  $\mathbf{A}^{(2)}$  can be distinguished via DE-GNN-1 with high probability. We formally state this result in the following corollary.

**Corollary 3.4.** Suppose two graphs are uniformly independently sampled from all  $n$ -sized  $r$ -regular graphs over  $V$  where  $3 \leq r < (2 \log n)^{1/2}$ . Then, within  $L \leq \lceil (\frac{1}{2} + \epsilon) \frac{\log n}{\log(r-1)} \rceil$  layers, DE-GNN-1 can distinguish these two graphs with probability  $1 - o(1)$  by being concatenated with injective set-pooling over all the representations of nodes.

One insightful observation is that the structural representation with a small set  $S$  may become easier to be learnt than the entire graph representation as the knowledge of the node set  $S$  can be viewed as a piece of side information. Models can be built to leverage such information, while when to compute the entire graph representation ( $S = V$ ), the model loses such side information. This could be a reason why the successful probability to learn structural representation of a small node set (Theorem 3.3) is higher than to that of an entire graph in (Corollary 3.4), though our derivation of the probabilistic bounds is not tight. Note that DE provides a convenient way to effectively capture such side information. One naive way to leverage the information of  $S$  is to simply annotate the nodes in  $S$  with binary encoding 1 and those out of  $S$  with 0. This is obviously a special case of DE but it is not as powerful as the general DE (even compared to the special case SPD). Think about setting the target node set in Eq.(1) as the entire graph ( $S = V$ ), where annotating the nodes in/out of  $S$  does not improve the representation power of WLGNN because all nodes are annotated as 1. However, DE still holds extra representation power: For example, we want to distinguish a graph with two disconnected 3-circles and a graph with a 6-circle. These two graphs generate different SPDs between nodes.

### 3.2.2 The Limitation of DE-GNN

Next, we show the limitation of DE-GNN. We prove that over a subclass of regular graphs, distance regular graphs (DRG), DE-1 is useless for structural representation. We provide the definition of DRG as follows while we refer interested readers to check more properties of DRGs in [35].

**Definition 3.5.** A *distance regular graph* is a regular graph such that for any two nodes  $v, u \in V$ , the number of vertices  $w$  s.t.  $\text{SPD}(w, v) = i$  and  $\text{SPD}(w, u) = j$ , only depends on  $i, j$  and  $\text{SPD}(v, u)$ .

The Shrikhande graph and the  $4 \times 4$  Rook's graph are two non-isomorphic DRGs shown in Fig. 2 (b) (We temporarily ignore the nodes colors which will be discussed later). For simplicity, we only consider connected DRGs that can be characterized by arrays of integers termed intersection arrays.

**Definition 3.6.** The *intersection array* of a connected DRG with diameter  $\Delta$  is an array of integers  $\{b_0, b_1, \dots, b_{\Delta-1}; c_1, c_2, \dots, c_{\Delta}\}$  such that for any node pair  $(u, v) \in V \times V$  that satisfies  $\text{SPD}(v, u) = j$ ,  $b_j$  is the number of nodes  $w$  that are neighbors of  $v$  and satisfy  $\text{SPD}(w, u) = j + 1$ , and  $c_j$  is the number of nodes  $w$  that are neighbors of  $v$  and satisfy  $\text{SPD}(w, u) = j - 1$ .

It is not hard to show that the two DRGs in Fig. 2 (b) share the same intersection array  $\{6, 3; 1, 2\}$ . The following theorem shows that over distance regular graphs, DE-GNN-1 requires discriminatory node/edge attributes to distinguish structures, which indicates the limitation of DE-1.

**Theorem 3.7.** Given any two nodes  $v, u \in V$ , consider two tuples  $\mathcal{T}_1 = (v, \mathbf{A}^{(1)})$  and  $\mathcal{T}_2 = (u, \mathbf{A}^{(2)})$  with graph structures  $A^{(1)}$  and  $A^{(2)}$  that correspond to two connected DRGs with a same intersection array. Then, DE-GNN-1 must use discriminatory node/edge attributes to distinguish  $\mathcal{T}_1$  and  $\mathcal{T}_2$ .

Note Theorem 3.7 only works for node representations using DE-1. Therefore, DE-GNN-1 may not associate distinguishable node representations in the two DRGs in Fig. 2 (b).

However, if we are to learn higher-order structural representations ( $|S| \geq 2$ ) with DE- $p$  ( $p \geq 2$ ), DE-GNN- $p$  may have even stronger representation power. We illustrate this point by considering structural representations of two node-pairs that form edges of the two DRGs respectively. Consider two node-pairs that correspond to two edges of these two graphs in Fig. 2 (b) respectively. Then, there exists a proper DE-GNN-2 via using SPD as DE-2, associating these two node-pairs with different representations. Moreover, by simply aggregating the obtained representations of all node-pairs into graph representations via a set-pooling, we may also distinguish these two graphs. Note that distinguishing the node-pairs of the two DRGs is really hard, because even the 2-WL test<sup>1</sup> will fail to distinguish any edges in the DRGs with a same intersection array and diameters exactly equal to 2<sup>2</sup>. This means that the recently proposed more powerful GNNs, such as RingGNN [30] and PPGN [27], will also fail in this case. However, it is possible to use DE-GNN-2 to distinguish those two DRGs.

It is interesting to generalize Theorem 3.3 to DRGs to demonstrate the power of DE-GNN- $p$  ( $p \geq 2$ ). However, missing analytic-friendly random models for DRGs makes such generalization challenging.

### 3.3 DEA-GNN— Distance Encoding-1’s as Controllers of the Message Aggregation

DE-GNN only uses DEs as initial node features. In this subsection, we further consider leveraging DE-1 between any two nodes to control the aggregation procedure of DE-GNN. Specifically, we propose DE-Aggregation-GNN (DEA-GNN) to do the following change

$$\text{AGG}(\{f_2(h_u^{(l)}, \mathbf{A}_{vu})\}_{u \in \mathcal{N}_v}) \rightarrow \text{AGG}(\{f_2(h_u^{(l)}, \mathbf{A}_{vu}), \zeta(u|v)\}_{u \in V}) \quad (4)$$

Note that the representation power of DEA-GNN is at least no worse than DE-GNN because the later one is specialized by aggregating the nodes with  $\zeta_{\text{spd}}(u|v) = 1$ , so Theorem 3.3, Corollary 3.4 are still true. Interestingly, its power is also limited by Theorem 3.7. We conclude as the follows.

**Corollary 3.8.** Theorem 3.3, Corollary 3.4 and Theorem 3.7 are still true for DEA-GNN.

The general form Eq. (4) that aggregates all nodes in each iteration holds more theoretical significance than practical usage due to scalability concern. In practice, the aggregation procedure of DEA-GNN may be trimmed by balancing the tradeoff between complexity and performance. For example, we may choose  $\zeta(u|v) = \zeta_{\text{spd}}(u|v)$ , and only aggregate the nodes  $u$  such that  $\zeta_{\text{spd}}(u|v) \leq K$ , *i.e.*,  $K$ -hop neighbors. Multi-hop aggregation allows avoiding the training issues of deep architecture, *e.g.*, gradient degeneration. Particularly, we may prove that  $K$ -hop aggregation decreases the number of layers  $L$  requested to  $\lceil (\frac{1}{2} + \epsilon) \frac{\log n}{K \log(r-1)} \rceil$  in Theorem 3.3 and Corollary 3.4 with proof in Appendix F. We may also choose  $\zeta(u|v) = \zeta_{\text{gpr}}(u|v)$  with non-negative  $\gamma_k$  in Eq. (3) and aggregate the nodes whose  $\zeta(u|v)$  are top- $K$  ranked among all  $u \in V$ . This manner is able to control fix-sized aggregation sets. As DEA-GNN does not show provably better representation power than DE-GNN, all the above approaches share the same theoretical power and limitations. However, in practice their specific performance may vary across datasets and applications.

<sup>1</sup>We follow the terminology 2-WL test in [39], which refines representations of node-pairs iteratively and is proved to be more powerful than 1-WL test. This 2-WL test is termed 2-WL’ test in [40] or 2-FWL test in [27]. A brief introduction of higher-order WL tests can be found in Appendix H.

<sup>2</sup>Actually, the 2-WL test may not distinguish edges in a special class of DRGs, termed strongly regular graphs (SRG) [39]. A connected SRG is a DRG with diameter constrained as 2 [41].

## 4 Related Work

Recently, extensive effort has been taken to improve the structural representation power of WL-GNN. From the theoretical perspective, most previous works only considered representations of entire graphs [26–30, 42, 43] while Srinivasan & Ribeiro initialized the investigation of structural representations of node sets [44] from the view of joint probabilistic distributions. Some works view GNNs as general approximators of invariant functions but the proposed models hold more theoretical implication than practical usage because of their dependence on polynomial( $n$ )-order tensors [29, 45, 46]. Ring-GNN [30] (or equivalently PPGN [27]), a relatively scalable model among them, was based on 3-order tensors and was proposed to achieve the expressive power of the 2-WL test (a brief introduction in Appendix H). However, Ring-GNN (PPGN) was proposed for entire-graph representations and cannot leverage the sparsity of the graph structure to be scalable enough to process large graphs [27, 30]. DE-GNN still benefits from such sparsity and are also used to represent node sets of arbitrary sizes. Moreover, our models theoretically behave orthogonal to Ring-GNN, as DE-2 can distinguish some non-isomorphic node-pairs that Ring-GNN fails to distinguish because the power of Ring-GNN is limited by the 2-WL test (Fig. 2 (b)).

Some works with empirical success inspire the proposal of DE, though we are the first one to derive theoretical characterization and leverage our theory to better those models as a return. SEAL [9] predicts links by reading out the representations of ego-networks of node-pairs. Although SEAL leverages a specific DE-2, the representations of those ego-networks are extracted via complex SortPooling [23]. However, we argue against such complex operations as DE-2 yields all the benefit of representations of node-pairs, as demonstrated by our experiments. PGNN [10] uses SPDs between each node and some anchor nodes to encode distance between nodes. As those encodings are not permutation invariant, PGNN holds worse inductive/generalization capability than our models.

Quite a few works targeted at revising neighbor aggregation procedure of WLGNN and thus are related to the DEA-GNN. However, none of them demystified their connection to DE or provided theoretical characterization. MixHop [47], PAGTN [31], MAT [17] essentially used  $\zeta_{spd}(u|v)$  to change the way of aggregation (Eq. (4)) while GDC [32] and PowerGNN [48] used variants of  $\zeta_{gpr}(u|v)$ . MixHop, GDC and PowerGNN are evaluated for node classification while PAGTN and MAT are evaluated for graph classification. GDC claims that the aggregation based on  $\zeta_{gpr}(u|v)$  does not help link prediction. However, we are able to show its empirical success for link prediction, as the key point missed by GDC is using DEs as extra node attributes (Appendix G.3). Note that as the above models are covered by DEA-GNN-1, their representation powers are all bounded by Theorem 3.7 according to Corollary 3.8.

## 5 Experiments

Extensive experiments<sup>3</sup> are conducted to evaluate our DE-GNN and DEA-GNN over three levels of tasks involving target node sets with sizes 1, 2 and 3 respectively: roles of nodes classification (Task 1), link prediction (Task 2), and triangle prediction (Task 3). Triangle prediction is to predict for any given subset of 3 nodes,  $\{u, v, w\}$ , whether links  $uv$ ,  $uw$ , and  $vw$  all exist. This task belongs to the more general class of higher-order network motif prediction tasks [49, 50] and has recently attracted much significance to [51–55]. We briefly introduce the experimental settings and save the details of the datasets and the model parameters to Appendix G.

**Dataset & Training.** We use the following six real graphs for the three tasks introduced above: Brazil-Airports (Task 1), Europe-Airports (1), USA-Airports (1), NS (2 & 3), PB (2), C.ele (2 & 3). For Task 1, the goal is to predict the passenger flow level of a given airport based solely on the flight traffic network. These airports datasets are chosen because the labels indicate the structural roles of nodes (4 levels in total from hubs to switches) rather than community identifiers of nodes as traditionally used [20, 21, 56]. For Tasks 2 & 3, the datasets were used by the strongest baseline [9], which consist of existing links/triangles plus the the same number of randomly sampled negative instances from those graphs. The positive test links/triangles are removed from the graphs during the training phase. For all tasks, we use 80%, 10%, 10% dataset splitting for training, validation, testing respectively. All the models are trained until loss converges and the testing performance of the best model on validation set is reported. We also report the experiments without validation sets that follow the original settings of the baselines [5, 9] in Appendix G.4.

<sup>3</sup>The code to evaluate our model can be downloaded from <https://github.com/snap-stanford/distance-encoding>.



Method	Data	Nodes (Task 1): Average Accuracy			Node-pairs (Task 2): AUC			Node-triads (Task 3): AUC	
		Bra.-Airports	Eur.-Airports	USA-Airports	C.elegans	NS	PB	C.elegans	NS
GCN [20]		64.55±4.18	54.83±2.69	56.58±1.11	74.03±0.99	74.21±1.72	89.78±0.99	80.94±0.51	81.72±1.50
SAGE [21]		70.65±5.33	56.29±3.21	50.85±2.83	73.91±0.32	79.96±1.40	90.23±0.74	84.72±0.40	84.06±1.14
GIN [16]		71.89±3.60†	57.05±4.08	58.87±2.12	75.58±0.59	87.75±0.56	91.11±0.52	86.42±1.12†	94.59±0.66†
Struc2vec [5]		70.88±4.26	57.94±4.01†	61.92±2.61†	72.11±0.31	82.76±0.59	90.47±0.60	77.72±0.58	81.93±0.61
PGNN [10]		N/A	N/A	N/A	78.20±0.33	94.88±0.77	89.72±0.32	86.36±0.74	79.36±1.49
SEAL [9]		N/A	N/A	N/A	88.26±0.56†	98.55±0.32†	94.18±0.57†	N/A	N/A
DE-GNN-SPD		<b>73.28±2.47</b>	56.98±2.79	<b>63.10±0.68*</b>	<b>89.37±0.17*</b>	<b>99.09±0.79</b>	<b>94.95±0.37*</b>	<b>92.17±0.72*</b>	<b>99.65±0.40*</b>
DE-GNN-LP		<b>75.10±3.80*</b>	58.41±3.20*	<b>64.16±1.70*</b>	86.27±0.33	98.01±0.55	91.45±0.41	86.24±0.18	<b>99.31±0.12*</b>
DEA-GNN-SPD		<b>75.37±3.25*</b>	57.99±2.39*	<b>63.28±1.59</b>	<b>90.05±0.26*</b>	<b>99.43±0.63*</b>	<b>94.49±0.24*</b>	<b>93.35±0.65*</b>	<b>99.84±0.14*</b>

Table 1: Performance in Average Accuracy and Area Under the ROC Curve (AUC) (mean in percentage  $\pm$  95% confidence level). † highlights the best baselines. \*, **bold font**, **bold font\*** respectively highlights the case where our models’ performance exceeds the best baseline on average, by 70% confidence, by 95% confidence.

**Baselines.** We choose six baselines. GCN [20], GraphSAGE(SAGE) [21], GIN [16] are representative methods of WLGNN. These models use node degrees as initial features when attributes are not available to keep inductive ability. Struc2vec [5] is a kernel-based method, particularly designed for structural representations of single nodes. PGNN [10] and SEAL [9] are also GNN-based methods: PGNN learns node positional embeddings and is not inductive for node classification; SEAL is particularly designed for link prediction by using entire-graph representations of ego-networks of node-pairs. SEAL outperforms other link prediction approaches such as VGAE [57]. The node initial features for these two models are set as the inductive setting suggested in their papers. We tune parameters of all baselines (Appendix G.5) and list their optimal performance here.

**Instantiation of DE-GNN and DEA-GNN.** We choose GCN as the basic WLGNN and implement three variants of DE-GNN over it. Note that GIN could be a more powerful basis while we tend to keep our models simple. The first two variants of Eq. (2) give us DE-GNN-SPD and DE-GNN-LP. The former uses SPD-based one-hot vectors  $\zeta_{sdp}$  as extra nodes attributes, and the latter uses the sequence of landing probabilities Eq. (2). Next, we consider an instantiation of Eq. (4), DEA-GNN-SPD that uses SPDs,  $\zeta(u|v) = \zeta_{sdp}(u|v) \leq K$ , to control the aggregation, which enables  $K$ -hop aggregation ( $K = 2, 3$  and the better performance will be used). DEA-GNN-SPD uses SPD-based one-hot vectors as extra nodes attributes. Appendix G.3 provides thorough discussion on implementation of the three variants and another implementation that uses Personalized PageRank scores to control the aggregation. Experiments are repeated 20 times using different seeds and we report the average.

**Results** are shown in Table 1. Regarding the node-level task, GIN outperforms other WLGNNs, which matches the theory in [16, 26, 29]. Struc2vec is also powerful though it is kernel-based. DE-GNN’s significantly outperform the baselines (except Eur.-Airport) which imply the power of DE-1’s. Among them, landing probabilities (LP) work slightly better than SPDs as DE-1’s.

Regarding node-pairs-level tasks, SEAL is the strongest baseline, as it is particularly designed for link prediction by using a special DE-2 plus a graph-level readout [23]. However, our DE-GNN-SPD performs even significantly better than SEAL: The numbers are close, but the difference is still significant; The decreases of error rates are always greater than 10% and achieve almost 30% over NS. This indicates that DE-2 is the key signal that makes SEAL work while the complex graph-level readout adopted by SEAL is not necessary. Moreover, our set-pooling form of DE-2 (Eq. (1)) decreases the dimension of DE-2 adopted in SEAL, which also betters the generalization of our models (See the detailed discussion in Appendix G.3). Moreover, for link prediction, SPD seems to be much better to be chosen as DE-2 than LP.

Regarding node-triads-level tasks, no baselines were particularly designed for this setting. We have not expected that GIN outperforms PGNN as PGNN captures node positional information that seems useful to predict triangles. We guess that the distortion of absolute positional embeddings learnt by PGNN may be the reason that limits its ability to distinguish structures with nodes in close positions: For example, the three nodes in a path of length two are close in position and the three nodes in a triangle are also close in position. However, this is not a problem for DE-3. We also conjecture that the gain based on DEs grows w.r.t. their orders (*i.e.*,  $|S|$  in Eq. (1)). Again, for triangle prediction, SPD seems to be much better to be chosen as DE-3 than LP.

Note that DEA-GNN-SPD further improves DE-GNN-SPD (by almost 1% across most of the tasks). This demonstrates the power of multi-hop aggregation (Eq. (4)). However, note that DEA-GNN-SPD needs to aggregate multi-hop neighbors simultaneously and thus pays an additional cost of scalability.

## Broader Impact

This work proposes a novel angle to systematically improve the structural representation power of GNNs. We break from the convention that previous works characterize and further improve the power of GNNs by intimating different-order WL tests [26, 27, 30, 58]. As far as we know, we are the first one to provide non-asymptotic analysis of the power of the proposed GNN models. Therefore, the proof techniques of Theorems 3.3, 3.7 may be expected to inspire new theoretical studies of GNNs and further better the practical usage of GNNs. Moreover, our models have good scalability by avoiding using the framework of WL tests, as higher-order WL tests are not able to leverage the sparsity of graphs. To be evaluated over extremely large graphs [59], our models can be simply trimmed and work on the ego-networks sampled with a limited size around the target node sets, just as the strategy adopted by GraphSAGE [21] and GraphSAINT [60]. **Therefore, our work may motivate practitioners to design and deploy more powerful GNNs in industrial pipelines to benefit the society.**

Distance encoding unifies the techniques of many GNN models [9, 17, 31, 32, 47] and provides a extremely general framework with clear theoretical characterization. In this paper, we only evaluate four specific instantiations over three levels of tasks. However, there are some other interesting instantiations and applications. For example, we expect a better usage of PageRank scores as edge attributes (Eq. (4)). Currently, our instantiation DEAGNN-PR simply uses those scores as weights in a weighted sum to aggregate node representations. We also have not considered any attention-based mechanism over DEs in aggregation while it seems to be useful [17, 22]. Researchers may try these directions in a more principled manner based on this work. Our approaches may also help other tasks based on structural representation learning, such as graph-level classification/regression [14, 16, 23, 26, 27] and subgraph counting [58], which correspond to **many applications with wide societal impact including drug discovery and structured data analysis.**

There are also two important implications coming from the observations of this work. First, Theorem 3.7 and Corollary 3.8 show the limitation of DE-1 over distance regular graphs, including the cases when DE-1's are used as node attributes or controllers of message aggregation. As distance regular graphs with the same intersection array have the important co-spectral property [35], we guess that DE-1 is a bridge to connect GNN frameworks to spectral approaches, two fundamental approaches in graph-structured data processing. This point sheds some light on the question left in [30] while more rigorous characterization is still needed. Second, as observed in the experiments, higher-order DE's induce larger gains as opposed to WLGNN, while Theorem 3.3 is not able to characterize this observation as the probability  $1 - o(\frac{1}{n})$  does not depend on the size  $p$ . We are sure that the probabilistic quantization in Theorem 3.3 is not tight, so it is interesting to see how such probability depends on  $p$  by deriving tighter bounds.

**We are not aware of any societal disadvantages of this research. Our experiments also do not leverage biases in the data.** We choose those datasets based on two rules that are irrelevant to ethics: (1) The labels for evaluation are graph-structured related; (2) Those datasets were used in some baselines so that we provide fair comparison. Regarding the point (1), this is the reason that we do not use the datasets, such as Cora, Citeseer and Pubmed [56], whose node labels that indicate the community belongings of nodes instead of the structural function of nodes. However, it is interesting to evaluate distance encoding techniques over those datasets with community-related labels in the future.

## Acknowledgements

The authors would like to thank Weihua Hu for raising the preliminary concept of distance encoding that initializes the investigation. The authors also would like to thank Jiaxuan You and Rex Ying for their insightful comments during the discussion. The authors also would thank Baharan Mirza-soleiman and Tailin Wu for their suggestions on the paper writing. The authors also would like to thank the NeurIPS reviewers for their insightful observations and actionable suggestions to improve the manuscript. This research has been supported in part by Purdue CS start-up, NSF CINES, NSF HDR, NSF Expeditions, NSF RAPID, DARPA MCS, DARPA ASED, ARO MURI, Stanford Data Science Initiative, Wu Tsai Neurosciences Institute, Chan Zuckerberg Biohub, Amazon, Boeing, Chase, Docomo, Hitachi, Huawei, JD.com, NVIDIA, Dell. J. L. is a Chan Zuckerberg Biohub investigator.

## References

- [1] W. L. Hamilton, R. Ying, and J. Leskovec, “Representation learning on graphs: Methods and applications,” *IEEE Data Engineering Bulletin*, vol. 40, no. 3, pp. 52–74, 2017.
- [2] S. P. Borgatti and M. G. Everett, “Notions of position in social network analysis,” *Sociological Methodology*, pp. 1–35, 1992.
- [3] K. Henderson, B. Gallagher, T. Eliassi-Rad, H. Tong, S. Basu, L. Akoglu, D. Koutra, C. Faloutsos, and L. Li, “Rolx: structural role extraction & mining in large graphs,” in *the ACM SIGKDD international conference on Knowledge discovery and data mining*, 2012, pp. 1231–1239.
- [4] R. A. Rossi and N. K. Ahmed, “Role discovery in networks,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 27, no. 4, pp. 1112–1131, 2014.
- [5] L. F. Ribeiro, P. H. Saverese, and D. R. Figueiredo, “struc2vec: Learning node representations from structural identity,” in *the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2017, pp. 385–394.
- [6] C. Donnat, M. Zitnik, D. Hallac, and J. Leskovec, “Learning structural node embeddings via diffusion wavelets,” in *the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018, pp. 1320–1329.
- [7] D. Liben-Nowell and J. Kleinberg, “The link-prediction problem for social networks,” *Journal of the American society for information science and technology*, vol. 58, no. 7, pp. 1019–1031, 2007.
- [8] M. Zhang and Y. Chen, “Weisfeiler-lehman neural machine for link prediction,” in *the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2017, pp. 575–583.
- [9] —, “Link prediction based on graph neural networks,” in *Advances in Neural Information Processing Systems*, 2018, pp. 5165–5175.
- [10] J. You, R. Ying, and J. Leskovec, “Position-aware graph neural networks,” in *International Conference on Machine Learning*, 2019, pp. 7134–7143.
- [11] N. Pržulj, “Biological network comparison using graphlet degree distribution,” *Bioinformatics*, vol. 23, no. 2, pp. 177–183, 2007.
- [12] L. A. Zager and G. C. Verghese, “Graph similarity scoring and matching,” *Applied mathematics letters*, vol. 21, no. 1, pp. 86–94, 2008.
- [13] N. Shervashidze, P. Schweitzer, E. J. v. Leeuwen, K. Mehlhorn, and K. M. Borgwardt, “Weisfeiler-lehman graph kernels,” *Journal of Machine Learning Research*, vol. 12, no. Sep, pp. 2539–2561, 2011.
- [14] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl, “Neural message passing for quantum chemistry,” in *International Conference on Machine Learning*, 2017, pp. 1263–1272.
- [15] Z. Ying, J. You, C. Morris, X. Ren, W. Hamilton, and J. Leskovec, “Hierarchical graph representation learning with differentiable pooling,” in *Advances in Neural Information Processing Systems*, 2018, pp. 4800–4810.
- [16] K. Xu, W. Hu, J. Leskovec, and S. Jegelka, “How powerful are graph neural networks?” in *International Conference on Learning Representations*, 2019.
- [17] Ł. Maziarka, T. Danel, S. Mucha, K. Rataj, J. Tabor, and S. Jastrzębski, “Molecule attention transformer,” *arXiv preprint arXiv:2002.08264*, 2020.
- [18] K. Hornik, M. Stinchcombe, H. White *et al.*, “Multilayer feedforward networks are universal approximators,” *Neural Networks*, vol. 2, no. 5, pp. 359–366, 1989.
- [19] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini, “The graph neural network model,” *IEEE Transactions on Neural Networks*, vol. 20, no. 1, pp. 61–80, 2008.
- [20] T. N. Kipf and M. Welling, “Semi-supervised classification with graph convolutional networks,” in *International Conference on Learning Representations*, 2017.
- [21] W. Hamilton, Z. Ying, and J. Leskovec, “Inductive representation learning on large graphs,” in *Advances in Neural Information Processing Systems*, 2017, pp. 1024–1034.

- [22] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, “Graph attention networks,” in *International Conference on Learning Representations*, 2018.
- [23] M. Zhang, Z. Cui, M. Neumann, and Y. Chen, “An end-to-end deep learning architecture for graph classification,” in *the AAAI Conference on Artificial Intelligence*, 2018, pp. 4438–4445.
- [24] P. W. Battaglia, J. B. Hamrick, V. Bapst, A. Sanchez-Gonzalez, V. Zambaldi, M. Malinowski, A. Tacchetti, D. Raposo, A. Santoro, R. Faulkner *et al.*, “Relational inductive biases, deep learning, and graph networks,” *arXiv preprint arXiv:1806.01261*, 2018.
- [25] B. Weisfeiler and A. Leman, “A reduction of a graph to a canonical form and an algebra arising during this reduction,” *Nauchno-Technicheskaya Informatsia*, 1968.
- [26] 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 *the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 4602–4609.
- [27] H. Maron, H. Ben-Hamu, H. Serviansky, and Y. Lipman, “Provably powerful graph networks,” in *Advances in Neural Information Processing Systems*, 2019, pp. 2153–2164.
- [28] R. Murphy, B. Srinivasan, V. Rao, and B. Ribeiro, “Relational pooling for graph representations,” in *International Conference on Machine Learning*, 2019.
- [29] H. Maron, H. Ben-Hamu, N. Shamir, and Y. Lipman, “Invariant and equivariant graph networks,” in *International Conference on Learning Representations*, 2019.
- [30] Z. Chen, S. Villar, L. Chen, and J. Bruna, “On the equivalence between graph isomorphism testing and function approximation with gnns,” in *Advances in Neural Information Processing Systems*, 2019, pp. 15 868–15 876.
- [31] B. Chen, R. Barzilay, and T. Jaakkola, “Path-augmented graph transformer network,” *ICML 2019 Workshop on Learning and Reasoning with Graph-Structured Data*, 2019.
- [32] J. Klicpera, S. Weissenberger, and S. Günnemann, “Diffusion improves graph learning,” in *Advances in Neural Information Processing Systems*, 2019, pp. 13 333–13 345.
- [33] E. Chien, J. Peng, P. Li, and O. Milenkovic, “Joint adaptive feature smoothing and topology extraction via generalized pagerank gnns,” *arXiv preprint arXiv:2006.07988*, 2020.
- [34] P. Li, I. Chien, and O. Milenkovic, “Optimizing generalized pagerank methods for seed-expansion community detection,” in *Advances in Neural Information Processing Systems*, 2019, pp. 11 705–11 716.
- [35] A. E. Brouwer and W. H. Haemers, “Distance-regular graphs,” in *Spectra of Graphs*. Springer, 2012, pp. 177–185.
- [36] L. Babai, “Graph isomorphism in quasipolynomial time,” in *Proceedings of the Forty-Eighth Annual ACM Symposium on Theory of Computing*, 2016, pp. 684–697.
- [37] A. Grover and J. Leskovec, “node2vec: Scalable feature learning for networks,” in *the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 855–864.
- [38] M. Zaheer, S. Kottur, S. Ravanbakhsh, B. Póczos, R. R. Salakhutdinov, and A. J. Smola, “Deep sets,” in *Advances in Neural Information Processing Systems*, 2017, pp. 3391–3401.
- [39] J.-Y. Cai, M. Fürer, and N. Immerman, “An optimal lower bound on the number of variables for graph identification,” *Combinatorica*, vol. 12, no. 4, pp. 389–410, 1992.
- [40] M. Grohe, *Descriptive complexity, canonisation, and definable graph structure theory*. Cambridge University Press, 2017, vol. 47.
- [41] A. E. Brouwer and W. H. Haemers, “Strongly regular graphs,” in *Spectra of Graphs*. Springer, 2012, pp. 115–149.
- [42] R. Kondor, H. T. Son, H. Pan, B. Anderson, and S. Trivedi, “Covariant compositional networks for learning graphs,” *arXiv preprint arXiv:1801.02144*, 2018.
- [43] R. Kondor and S. Trivedi, “On the generalization of equivariance and convolution in neural networks to the action of compact groups,” in *International Conference on Machine Learning*, 2018, pp. 2747–2755.
- [44] B. Srinivasan and B. Ribeiro, “On the equivalence between node embeddings and structural graph representations,” in *International Conference on Learning Representations*, 2020.

- [45] H. Maron, E. Fetaya, N. Segol, and Y. Lipman, “On the universality of invariant networks,” in *International Conference on Machine Learning*, 2019, pp. 4363–4371.
- [46] N. Keriven and G. Peyré, “Universal invariant and equivariant graph neural networks,” in *Advances in Neural Information Processing Systems*, 2019, pp. 7090–7099.
- [47] S. Abu-El-Haija, B. Perozzi, A. Kapoor, N. Alipourfard, K. Lerman, H. Harutyunyan, G. Ver Steeg, and A. Galstyan, “Mixhop: Higher-order graph convolutional architectures via sparsified neighborhood mixing,” in *International Conference on Machine Learning*, 2019, pp. 21–29.
- [48] Z. Chen, L. Li, and J. Bruna, “Supervised community detection with line graph neural networks,” in *International Conference on Learning Representations*, 2019.
- [49] A. R. Benson, R. Abebe, M. T. Schaub, A. Jadbabaie, and J. Kleinberg, “Simplicial closure and higher-order link prediction,” *the Proceedings of the National Academy of Sciences*, vol. 115, no. 48, pp. 11 221–11 230, 2018.
- [50] H. Nassar, A. R. Benson, and D. F. Gleich, “Pairwise link prediction,” in *the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 2019, pp. 386–393.
- [51] A. R. Benson, D. F. Gleich, and J. Leskovec, “Higher-order organization of complex networks,” *Science*, vol. 353, no. 6295, pp. 163–166, 2016.
- [52] P. Li and O. Milenkovic, “Inhomogeneous hypergraph clustering with applications,” in *Advances in Neural Information Processing Systems*, 2017, pp. 2308–2318.
- [53] P. Li, G. J. Puleo, and O. Milenkovic, “Motif and hypergraph correlation clustering,” *IEEE Transactions on Information Theory*, 2019.
- [54] H. Yin, A. R. Benson, J. Leskovec, and D. F. Gleich, “Local higher-order graph clustering,” in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2017, pp. 555–564.
- [55] C. E. Tsourakakis, J. Pachocki, and M. Mitzenmacher, “Scalable motif-aware graph clustering,” in *Proceedings of the 26th International Conference on World Wide Web*, 2017, pp. 1451–1460.
- [56] P. Sen, G. Namata, M. Bilgic, L. Getoor, B. Galligher, and T. Eliassi-Rad, “Collective classification in network data,” *AI magazine*, vol. 29, no. 3, pp. 93–93, 2008.
- [57] T. N. Kipf and M. Welling, “Variational graph auto-encoders,” *NeurIPS Bayesian Deep Learning Workshop*, 2016.
- [58] Z. Chen, L. Chen, S. Villar, and J. Bruna, “Can graph neural networks count substructures?” *arXiv preprint arXiv:2002.04025*, 2020.
- [59] W. Hu, M. Fey, M. Zitnik, Y. Dong, H. Ren, B. Liu, M. Catasta, and J. Leskovec, “Open graph benchmark: Datasets for machine learning on graphs,” *arXiv preprint arXiv:2005.00687*, 2020.
- [60] H. Zeng, H. Zhou, A. Srivastava, R. Kannan, and V. Prasanna, “Graphsaint: Graph sampling based inductive learning method,” in *International Conference on Learning Representations*, 2020.
- [61] B. Bollobás, “A probabilistic proof of an asymptotic formula for the number of labelled regular graphs,” *European Journal of Combinatorics*, vol. 1, no. 4, pp. 311–316, 1980.
- [62] V. Arvind, F. Fuhlbrück, J. Köbler, and O. Verbitsky, “On weisfeiler-leman invariance: subgraph counts and related graph properties,” in *International Symposium on Fundamentals of Computation Theory*. Springer, 2019, pp. 111–125.
- [63] L. W. Beineke and R. J. Wilson, Eds., *Graph Connections: Relationships between Graph Theory and Other Areas of Mathematics*. Oxford University Press, 1997.
- [64] R. Ackland *et al.*, “Mapping the us political blogosphere: Are conservative bloggers more prominent?” in *BlogTalk Downunder 2005 Conference, Sydney*. BlogTalk Downunder 2005 Conference, Sydney, 2005.
- [65] C. C. Kaiser, Marcus; Hilgetag, “Nonoptimal component placement, but short processing paths, due to long-distance projections in neural systems,” *PLoS Computational Biology*, vol. 2, no. 7, p. 95, 2006.

- [66] M. E. Newman, “Finding community structure in networks using the eigenvectors of matrices,” *Physical review E*, vol. 74, no. 3, p. 036104, 2006.
- [67] G. Jeh and J. Widom, “Scaling personalized web search,” in *the International Conference on World Wide Web*, 2003, pp. 271–279.
- [68] F. Chung, “The heat kernel as the pagerank of a graph,” *Proceedings of the National Academy of Sciences*, vol. 104, no. 50, pp. 19 735–19 740, 2007.
- [69] D. F. Gleich and R. A. Rossi, “A dynamical system for pagerank with time-dependent teleportation,” *Internet Mathematics*, vol. 10, no. 1-2, pp. 188–217, 2014.

## Appendix

### A Proof of Universal Approximate Theorem for Structural Representation

We restate Theorem 2.6: If the structural representation  $\Gamma$  can distinguish any two non-isomorphic tuples  $\mathcal{T}^{(1)}$  and  $\mathcal{T}^{(2)}$  in  $\Omega_p$ , then for any invariant function  $f : \Omega_p \rightarrow \mathbb{R}$ ,  $f$  can be universally approximated by  $\Gamma$  via a 3-layer feed-forward neural network with ReLU as rectifiers, as long as

- The feature space  $\mathcal{A}$  is compact.
- $f(S, \cdot)$  is continuous over  $\mathcal{A}$  for any  $S \in \mathcal{P}_p(V)$ .

*Proof.* This result is a direct generalization of Theorem 4 [30]. Specifically, we extend the statement of representing graphs featured by  $\mathbf{A}$  to that of representing structures featured by  $(S, \mathbf{A})$ .

Recall the original space  $\mathcal{A} \subset \mathbb{R}^{n \times n \times k}$ . We define a space  $\mathcal{A}' \subset \mathbb{R}^{n \times n \times (k+1)}$ : For any  $\mathbf{A}' \in \mathcal{A}'$ , its slice in the first  $k$  dimensions of 3-rd mode, i.e.,  $\mathbf{A}'_{\cdot, \cdot, 1:k}$ , is in  $\mathcal{A}$  and the slice corresponds to the last dimension of 3-rd mode, i.e.,  $\mathbf{A}'_{\cdot, \cdot, k+1}$  is a diagonal matrix where the diagonal components could be only 0 or 1. Then, we may build a bijective mapping between  $\mathbf{A}' \in \mathcal{A}'$  and  $(S, \mathbf{A}) \in \Omega_p$  by

$$\mathbf{A}'_{\cdot, \cdot, 1:k} = \mathbf{A}, \quad \mathbf{A}'_{u, u, k+1} = 1 \text{ if } u \in S \text{ or } 0 \text{ if } u \notin S$$

As  $\mathcal{A}$  is compact in  $\mathbb{R}^{n \times n \times k}$  and we have only finite possible choices of  $\mathbf{A}'_{\cdot, \cdot, k+1}$ , actually  $\binom{n}{|S|}$ , the space  $\mathcal{A}'$  is compact in  $\mathbb{R}^{n \times n \times (k+1)}$ .

Then, we may transfer all definitions from  $\Omega_p$  to  $\mathcal{A}'$ . Specifically, the structural representation  $\Gamma$  that distinguishes any two non-isomorphic tuples  $\mathcal{T}^{(1)}$  and  $\mathcal{T}^{(2)}$  in  $\Omega_p$  defines  $\Gamma' : \mathcal{A}' \rightarrow \mathbb{R}^d$  that distinguishes any two non-isomorphic tensors  $\mathbf{A}'^{(1)}$  and  $\mathbf{A}'^{(2)}$  in  $\mathcal{A}'$ , as  $\mathcal{T}^{(i)}$  and  $\mathbf{A}'^{(i)}$  form a bijective mapping for  $i = 1, 2$ . Moreover, one invariant function  $f : \Omega_p \rightarrow \mathbb{R}$  also defines another invariant function  $f' : \mathcal{A}' \rightarrow \mathbb{R}$ , as  $\mathcal{T}^{(i)}$  and  $\mathbf{A}'^{(i)}$  form a bijective mapping for  $i = 1, 2$ .

Suppose the original metric over  $\mathcal{A}$  is denoted by  $\mathcal{M} : \mathcal{A} \times \mathcal{A} \rightarrow \mathbb{R}_{\geq 0}$ . Define a metric over  $\mathcal{A}'$  as

$$\mathcal{M}'(\mathbf{A}'^{(1)}, \mathbf{A}'^{(2)}) = \mathcal{M}(\mathbf{A}'^{(1)}_{\cdot, \cdot, 1:k}, \mathbf{A}'^{(2)}_{\cdot, \cdot, 1:k}) + \sum_{u \in V} \mathbb{1}_{\mathbf{A}'^{(1)}_{u, u, k+1} \neq \mathbf{A}'^{(2)}_{u, u, k+1}}.$$

Then, it is easy to show that we have the following lemma based on the definition of continuity.

**Lemma A.1.** If  $f(S, \cdot)$  is continuous over  $\mathcal{A}$  for any  $S \in \mathcal{P}_p(V)$  with respect to  $\mathcal{M}$ , then  $f'$  is continuous over  $\mathcal{A}'$  with respect to  $\mathcal{M}'$ .

Now, we only need to use Theorem 4 [30] to prove the statement. Actually the dimensions of  $\Gamma'$  overall forms a collection of one-dimensional functions  $\Xi = (\Gamma'[i])_{i \in [d]}$ , where  $\Gamma'[i]$  is the  $i$ th component of  $\Gamma' \in \mathbb{R}^d$ . According to the definition of  $\Gamma'$ , we know  $\Xi$  distinguishes all the non-isomorphic  $\mathbf{A}'^{(1)}$  and  $\mathbf{A}'^{(2)}$  in  $\mathcal{A}'$ . Moreover, because  $\mathcal{A}'$  is compact and  $f'$  is continuous over  $\mathcal{A}'$ , Theorem 4 [30] shows that the arbitrary invariant function  $f'$  defined on  $\mathcal{A}'$  can be universally approximated by  $\Xi$  via 3-layer feed-forward neural networks with ReLU as rectifiers. Recall  $f$  and  $f'$  are bijective, and  $\Gamma$ ,  $\Gamma'$ , and  $\Xi$  are mutually bijective. Therefore, we claim that  $f$  can be universally approximated by  $\Gamma$  via 3-layer feed-forward neural networks with ReLU as rectifiers.  $\square$

### B Proof for The Power of WLGNN for Structural Representation

We restate Theorem 2.7: Consider two tuples  $\mathcal{T}_1 = (S^{(1)}, \mathbf{A}^{(1)})$  and  $\mathcal{T}_2 = (S^{(2)}, \mathbf{A}^{(2)})$  in  $\Omega_p$ . If  $\mathcal{T}_1, \mathcal{T}_2$  cannot be distinguished by the 1-WL test, then the corresponding outputs of WLGNN satisfy  $\Gamma(\mathcal{T}_1) = \Gamma(\mathcal{T}_2)$ . On the other side, if they can be distinguished by the 1-WL test and suppose aggregation operations (AGG) and feed-forward neural networks  $f_1, f_2$  are all injective mappings, then a large enough number of layers  $L$ , the outputs of WLGNN satisfy  $\Gamma(\mathcal{T}_1) \neq \Gamma(\mathcal{T}_2)$ .

*Proof.* There is no fundamental difficulty to generalize the results from the case of graph representation to that of structural representation, because the only difference according to WLGNN is the final

readout step (AGG( $\cdot$ )), which works on a subset of nodes instead of the entire node set. Therefore, the same logic of proofs of Lemma 2 and Theorem 3 in [16] can be directly applied for structural representation learning with little revision.  $\square$

### C Proof for The Power of DE — Theorem 3.3

We restate Theorem 3.3: Given two fixed-sized sets  $S^{(1)}, S^{(2)} \subset V$ ,  $|S^{(1)}| = |S^{(2)}| = p$ . Consider two tuples  $\mathcal{T}^{(1)} = (S^{(1)}, \mathbf{A}^{(1)})$  and  $\mathcal{T}^{(2)} = (S^{(2)}, \mathbf{A}^{(2)})$  in the most difficult setting when features  $\mathbf{A}^{(1)}$  and  $\mathbf{A}^{(2)}$  are only different in graph structures specified by  $A^{(1)}$  and  $A^{(2)}$  respectively. Suppose  $A^{(1)}$  and  $A^{(2)}$  are uniformly independently sampled from all  $r$ -regular graphs over  $V$  where  $3 \leq r < (2 \log n)^{1/2}$ . Then, for any constant  $\epsilon > 0$ , there exist a proper DE-GNN- $p$  with layers  $L < (\frac{1}{2} + \epsilon) \frac{\log n}{\log(r-1)}$ , using DE- $p$   $\zeta(u|S^{(1)})$ ,  $\zeta(u|S^{(2)})$  for all  $u \in V$  such that with probability  $1 - o(n^{-1})$ , its outputs  $\Gamma(\mathcal{T}^{(1)}) \neq \Gamma(\mathcal{T}^{(2)})$ . Specifically,  $f_3$  can be simply chosen as SPD, *i.e.*,  $\zeta(u|v) = \zeta_{spd}(u|v)$ . The big-O notation is with respect to  $n$ .

*Proof.* To prove the statement, we only need to prove the case that  $|S^{(1)}| = |S^{(2)}| = 1$ ,  $\zeta(u|v) = \zeta_{spd}(u|v)$  because of the following lemma.

**Lemma C.1.** Suppose the statement is true when  $|S^{(1)}| = |S^{(2)}| = 1$ ,  $\zeta(u|v) = \zeta_{spd}(u|v)$ . Then, the statement is also true for the case when  $|S^{(1)}| = |S^{(2)}| = p > 1$  for some fixed  $p$ , and  $\zeta(u|v)$  is a neural network fed with the list of landing probabilities.

*Proof.* We first focus on the case when DE is chosen as SPD, *i.e.*,  $\zeta(u|v) = \zeta_{spd}(u|v)$ . We want to use the results of the case  $|S^{(1)}| = |S^{(2)}| = 1$  to prove that of the case  $|S^{(1)}| = |S^{(2)}| > 1$ . Suppose  $|S^{(1)}| = |S^{(2)}| > 1$ . We choose an arbitrary node from  $S^{(1)}$ , say  $w_1$ . As we assume the statement is true for the single node case, for any node in  $S^{(2)}$ , say  $w_2$ , DE-GNN with DEs  $\zeta_{spd}(u|w_1)$ ,  $\zeta_{spd}(u|w_2)$  is able to distinguish two tuples  $(w_1, \mathbf{A}^{(1)})$  and  $(w_2, \mathbf{A}^{(2)})$ , with probability at least  $1 - o(n^{-1})$ . Given that the space of SPD is countable and AGG in Eq. (1) is injective,  $\zeta_{spd}(u|S^{(1)})$  and  $\zeta_{spd}(u|S^{(2)})$  are different if  $\zeta_{spd}(u|w_1)$  is different from any  $\zeta_{spd}(u|w_2)$  ( $w_2 \in S^{(2)}$ ), which happens with probability at least  $1 - |S^{(2)}|o(n^{-1}) = 1 - o(n^{-1})$ . Therefore, DE-GNN with DEs  $\zeta_{spd}(u|S^{(1)})$ ,  $\zeta_{spd}(u|S^{(2)})$  is also able to distinguish two tuples  $(w_1, \mathbf{A}^{(1)})$  and  $(w_2, \mathbf{A}^{(2)})$ , with probability at least  $1 - o(n^{-1})$ . Based on the union bound, we know that DE-GNN with DEs  $\zeta_{spd}(u|S^{(1)})$ ,  $\zeta_{spd}(u|S^{(2)})$  is able to distinguish two tuples  $(w_1, \mathbf{A}^{(1)})$  and  $(v, \mathbf{A}^{(2)})$  for any  $v \in S^{(2)}$ , with probability at least  $1 - |S^{(2)}|o(n^{-1}) = 1 - o(n^{-1})$ . Therefore, we prove the capability to generalize the result from the single node case to the multiple node cases.

Now, let us generalize the result from  $\zeta_{spd}(u|v)$  to arbitrary  $\zeta(u|v)$  represented by neural networks fed with the list of landing probabilities. As  $\zeta_{spd}(u|v)$  is indeed a function of the list of landing probabilities (Eq. (2)), the general  $\zeta(u|v)$  should have stronger discriminatory power unless neural networks cannot provide a good mapping from the list of landing probabilities to SPD. However, we do not have to worry this because the list of landing probabilities fortunately lies a countable space for unweighted graphs: 1) The dimension of this list is countable (finite in practice); 2) Each component of this list is always a rational number if the graph is unweighted. According to our assumption,  $f_3$  is allowed to have an injective mapping over the list of landing probabilities. Therefore, neural networks on the list of landing probabilities will not decrease the representation power that is just based on  $\zeta_{spd}(u|v)$ .  $\square$

Keep in mind that Lemma C.1 could be loose. We leave the tight probability bound for the  $p > 2$  case in the future.

**Outline:** From now on, we focus on the single node case, *i.e.*  $|S^{(1)}| = |S^{(2)}| = 1$ , with SPD as the DE-1 ( $f_3$ ), *i.e.*,  $\zeta(u|v) = \zeta_{spd}(u|v)$ . Without loss of generality, we suppose  $S^{(1)} = S^{(2)} = \{u\}$ . As SPD is countable, there exists a proper DE-GNN that guarantees that all the operations are injective, which follows the basic condition used in [16]. Because of the iterative procedure of DE-GNN and all mappings are injection, the label of node  $u$  only depends on the subtree with depth  $L$  rooted at  $u$  (See the illustration of the subtree rooted at a given node in Fig. 3). Recall  $L$  is the number of layers



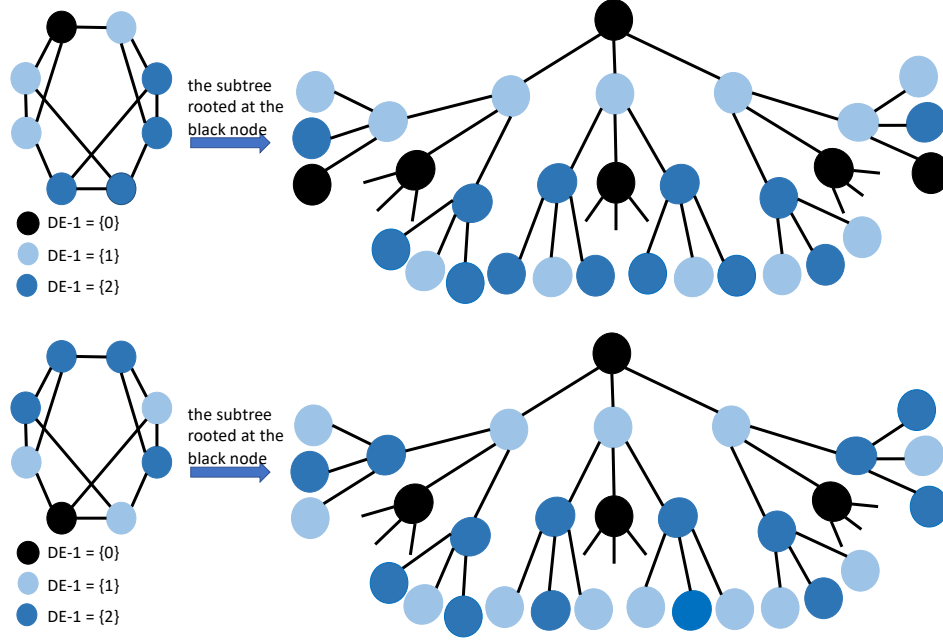


Figure 3: The subtree rooted at a node: In the left two graphs, the black nodes are the target nodes who structural representations are to be learnt. Different colors of the nodes correspond to different SPDs with respect to the target nodes. The right two trees correspond to the subtrees rooted at these two target nodes respectively. DE-GNN essentially works from bottom to top along these subtrees to obtain the representations of the target nodes. Different types of subtrees yield different representations based on a proper DE-GNN. In this example, these two target nodes are all from 3-regular graphs so they cannot be distinguished via WLGNN without DE-1’s or the informative node/edge attributes. However, with DE-1’s, we can see the corresponding subtrees of these two nodes are different and the difference appears in the second layer.

in DE-GNN. Therefore, we only need to show that  $A^{(1)}$  and  $A^{(2)}$ , which are uniformly sampled  $r$ -regular graphs, with probability at most  $o(n^{-1})$ , have the same subtrees rooted at  $u$  given SPDs as initial node labels. To show this, our proof contains four steps. Note that all through the following proof, we assume that  $n$  is very large and  $\epsilon$  is any small positive constant that is independent from  $n$ .

- We first explain that we are able to work on the configuration model of  $r$ -regular graphs proposed in [61] which associates uniform measure over all  $r$ -regular graphs. Given the condition  $r < (2 \log n)^{1/2}$ , there are a large portion ( $\Omega(n^{-1/2})$ ) of all the graphs generated by this model are simple (without self-loops and multi-edges)  $r$ -regular graphs [61]. Since the configuration model alleviates the difficulty to analyze the dependence between edges in  $r$ -regular graphs, we consider the graphs generated by the configuration model for the next two steps.
- Suppose the set of nodes that are associated with SPD=  $k$  from  $u$  is denoted by  $Q_k$ , and the number of edges that connect the nodes in  $Q_k$  and those in  $Q_{k+1}$  is denoted by  $p_k$ . We prove that with probability  $1 - o(n^{-\frac{3}{2}})$ , for all  $k \in (\frac{\epsilon}{5} \frac{\log n}{\log(r-1)} + 1, (\frac{2}{3} - \epsilon) \frac{\log n}{\log(r-1)})$ ,  $|Q_k| \geq (r - 1 - \epsilon)^{k-1}$  and  $p_k \geq (r - 1 - \epsilon)|Q_k|$  based on the configuration model.
- Next, we define the edge configuration between  $Q_k$  and  $Q_{k+1}$  as a list  $C_k = (a_{1,k}, a_{2,k}, \dots)$  where  $a_{i,k}$  denotes the number of nodes in  $Q_{k+1}$  of which each has exactly  $i$  edges from  $Q_k$ . We prove that for each  $k \in (\frac{1}{2} \frac{\log n}{\log(r-1-\epsilon)}, \frac{4}{7} \frac{\log n}{\log(r-1-\epsilon)})$ , as the edges between  $Q_k$  and  $Q_{k+1}$  are so many, there are too much randomness that makes each type of edge configuration  $C_k$  appear with only limited probability  $\mathbb{P}(C_k) = O(\frac{n^{1/2}}{p_k})$ . Recall that  $p_k$  is defined in step 2. Then, given any  $\epsilon \frac{\log n}{\log(r-1-\epsilon)}$  many  $k$ ’s, the probability that  $A^{(1)}$  and  $A^{(2)}$  have all

the same edge configurations for these  $k$ 's is bounded by  $\prod_k \mathbb{P}(C_k) \sim o(n^{-\frac{3}{2}})$ . Therefore, we only need to consider edge configurations for  $k \in (\frac{1}{2} \frac{\log n}{\log(r-1-\epsilon)}, (\frac{1}{2} + \epsilon) \frac{\log n}{\log(r-1-\epsilon)})$  to distinguish  $A^{(1)}$  and  $A^{(2)}$ , as this will give us  $1 - o(n^{-3/2})$  probability to succeed.

- Since there are at least  $\Omega(n^{-1/2})$  of all the graphs generated by the configuration model that are simple  $r$ -regular graphs, and there are at most  $o(n^{-\frac{3}{2}})$  probability that  $A^{(1)}$  and  $A^{(2)}$  share the same subtrees rooted at  $u$ , there are at most  $o(n^{-\frac{3}{2}}/n^{-1/2}) = o(n^{-1})$  probability that  $A^{(1)}$  and  $A^{(2)}$  are simple  $r$ -regular graphs and share the same subtrees rooted at  $u$ , which concludes the proof.

**Step 1:** We first introduce the configuration model proposed in [61] for  $r$ -regular graphs of  $n$  nodes. Suppose we have  $n$  sets of items,  $W_u, u \in [n]$ , where each set corresponds to one node in  $[n]$ . Each set  $W_u$  has  $r$  items. Now, we randomly partition all these  $nr$  items into  $\frac{nr}{2}$  pairs. Then, each partitioning result corresponds to a  $r$ -regular graph: if a pair contains items from  $W_u$  and  $W_v$ , then there is an edge between nodes  $u$  and  $v$  in the graph. Note that such partitioning results may render self-loops and multi-edges. Of course, we would like to consider only simple graphs which do not have self-loops and multi-edges. For this, the theory in [61] shows that for all these  $r$ -regular graphs, if  $r < (2 \log n)^{1/2}$ , there are about  $\exp(-\frac{r^2-1}{4})$  portion among them, *i.e.*,  $\Omega(n^{-1/2})$ , which are simple graphs.

**Step 2:** Now, we consider a graph that is uniformly sampled from the configuration model. Recall that the set of nodes that are associated with  $SPD = k$  from  $u$  is denoted by  $Q_k$ , and the number of edges that connect the nodes in  $Q_k$  and those in  $Q_{k+1}$  is denoted by  $p_k$ . Now, we prove that there exists a small constant  $\epsilon > 0$ , such that with probability  $1 - o(n^{-\frac{3}{2}})$ , for all  $k \in (\frac{\epsilon}{5} \frac{\log n}{\log(r-1)} + 1, (\frac{2}{3} - \epsilon) \frac{\log n}{\log(r-1)})$ ,  $|Q_k| \geq (r-1-\epsilon)^{k-1}$  and  $p_k \geq (r-1-\epsilon)|Q_k|$ . We prove an even stronger lemma that gives the previous argument via a union bound and doing induction over all  $k \in (\frac{\epsilon}{5} \frac{\log n}{\log(r-1)} + 1, (\frac{2}{3} - \epsilon) \frac{\log n}{\log(r-1)})$ .

**Lemma C.2.** There exists a small constant  $\epsilon > 0$ , with probability  $1 - O(n^{-2+\epsilon})$ , such that: 1) For any  $k < (\frac{2}{3} - \epsilon) \frac{\log n}{\log(r-1)}$ , if  $|Q_k| \geq n^{\epsilon/5}$ ,  $|Q_{k+1}| \geq p_k - |Q_k|^{1/2}$  and  $p_k \geq (r-1)|Q_k| - |Q_k|^{1/2}$ ; 2) When  $k = \lceil \frac{\epsilon}{5} \frac{\log n}{\log(r-1)} \rceil + 1$ ,  $|Q_k| \geq (r-1)^{k-1} = n^{\epsilon/5}$ .

*Proof.* We consider the following procedure to generate the graph based on the configuration model. Recall the node  $u$  is the targeted node, or the node with  $SPD = 0$  to the target node. We start from generating the edges attached to this node. We start from the set  $W_u$  and generate the  $r$  pairs with at least one item in  $W_u$ . Then, we have all the nodes in  $Q_1$ . Based on the set  $\cup_{v \in Q_1} W_v$ , we generate all the  $(r-1)|Q_1|$  pairs with at least one item in  $\cup_{v \in Q_1} W_v$ , and we have all the nodes in  $Q_2$ . The procedure goes on so on and so forth, from  $Q_k$  to  $Q_{k+1}$ .

Now, we prove 1). First, we prepare some inequalities. We have  $|Q_k| \leq r(r-1)^{k-1} < n^{2/3-\epsilon}$  by the assumption on  $k$ . For  $i \geq \lceil |Q_k|^{1/2} \rceil$ , we have

$$\frac{|Q_k| \cdot (r-1)|Q_k|}{n \cdot i} \leq n^{-\epsilon} \quad (5)$$

Moreover, Recall  $|Q_k| < n^{2/3-\epsilon}$ . As  $|Q_k| \geq n^{\epsilon/5}$ , then

$$\left( \frac{e(r-1)|Q_k|^{3/2}}{n} \right)^{\lceil |Q_k|^{1/2} \rceil} = O(n^{-2+\epsilon}). \quad (6)$$

This inequality is very crude.

Now, we go back to prove the bound for  $p_k$ . Recall the definition of  $p_k$  that is the number of edges between  $Q_k$  and  $Q_{k+1}$ . Then, the number of edges that are generated with both end-nodes are in  $Q_k$  is  $(r-1)|Q_k| - p_k$ . As we suppose the edges are generated sequentially, the probability to generate an edge whose two end-nodes are in  $Q_k$  is upper bounded by  $\frac{(r-1)|Q_k|}{r(n - \sum_{j=0}^k |Q_j|)} < \frac{|Q_k|}{n}$  where we use

$\sum_{j=0}^k |Q_j| \leq r(r-1)^k = O(n^{2/3})$ . Then, the probability that  $(r-1)|Q_k| - p_k > |Q_k|^{1/2}$  is upper bounded by (just summing over all possible  $(r-1)|Q_k| - p_k = i > |Q_k|^{1/2}$ )

$$\begin{aligned} \sum_{i=\lceil |Q_k|^{1/2} \rceil}^{(r-1)|Q_k|} \left[ \frac{|Q_k|}{n} \right]^i \binom{(r-1)|Q_k|}{i} &\stackrel{(5)}{<} \left[ \frac{|Q_k|}{n} \right]^{\lceil |Q_k|^{1/2} \rceil} \binom{(r-1)|Q_k|}{\lceil |Q_k|^{1/2} \rceil} \sum_{i \geq 0} n^{-i\epsilon} \\ &< c_1 \left[ \frac{|Q_k|}{n} \right]^{\lceil |Q_k|^{1/2} \rceil} \binom{(r-1)|Q_k|}{\lceil |Q_k|^{1/2} \rceil} \\ &< c_2 \left[ \frac{|Q_k|}{n} \right]^{\lceil |Q_k|^{1/2} \rceil} \left[ \frac{e(r-1)|Q_k|}{\lceil |Q_k|^{1/2} \rceil} \right]^{\lceil |Q_k|^{1/2} \rceil} \stackrel{(6)}{=} O(n^{-2+\epsilon}) \end{aligned}$$

where  $c_1, c_2$  are constants, the numbers above the equality/inequality signs refer to which equations are used.

Next, we prove the bound for  $|Q_{k+1}| \geq p_k - |Q_k|^{1/2}$ . Again, if the edges are generated sequentially,  $p_k - Q_{k+1}$  indicates the number of edges whose end-nodes in  $Q_{k+1}$  also belong to other edges that has been generated between  $Q_k$  and  $Q_{k+1}$ . The probability of this edge is upper bounded by  $\frac{|Q_{k+1}|}{r(n - \sum_{j=0}^{k+1} |Q_j|)} < \frac{(r-1)|Q_k|}{r(n - \sum_{j=0}^{k+1} |Q_j|)} \leq \frac{|Q_k|}{n}$ . Then, the probability that  $p_k - Q_{k+1} > |Q_k|^{1/2}$  is again upper bounded by (just summing over all possible  $p_k - Q_{k+1} = i > |Q_k|^{1/2}$ )

$$\sum_{i=\lceil |Q_k|^{1/2} \rceil}^{(r-1)|Q_k|} \left[ \frac{|Q_k|}{n} \right]^i \binom{(r-1)|Q_k|}{i} = O(n^{-2+\epsilon})$$

Till now, we have proved the statement 1).

Now, we prove the statement 2). Actually, at the time when  $Q_k$  is generated, the number of edges having been generated is at most  $r(r-1)^k - 2$ . These edges cover at most  $r(r-1)^k - 1$  nodes. When  $k = \lceil \frac{\epsilon}{5} \frac{\log n}{\log(r-1)} \rceil + 1$ , we claim that with probability  $1 - O(n^{-2+\epsilon})$ , at most 1 edge among these edges when generated is not connected to a new node. This is because if there are more than 1 such edges, the probability is at most (by summing over  $i$  such edges)

$$\sum_{i=2}^{r(r-1)^k} \left[ \frac{r(r-1)^k}{n - r(r-1)^k} \right]^i \binom{r(r-1)^k}{i} \leq c_3 \left( \frac{r^2 n^{\epsilon/5}}{n} \right)^2 \binom{r^2 n^{\epsilon/5}}{2} = O(n^{-2+\epsilon}).$$

We use this result to give a lower bound of  $|Q_k|$  for  $k = \lceil \frac{\epsilon}{5} \frac{\log n}{\log(r-1)} \rceil + 1$ . Because there is at most 1 edge when generated do not connect to a new node. The worst case appears when two items in  $W_u$  are mutually connected, which leads to  $|Q_1| \geq r-2 \geq 1$ . All edges after  $Q_1$  is generated are connected to new nodes and furthermore  $|Q_k| \geq (r-1)^{k-1} \geq n^{\epsilon/5}$ .  $\square$

**Step 3:** We start to consider the edge configuration between  $Q_k$  and  $Q_{k+1}$  for  $k \in (\frac{1}{2} \frac{\log n}{\log(r-1-\epsilon)}, \frac{4}{7} \frac{\log n}{\log(r-1-\epsilon)})$ . We focus our attention on the graphs that satisfy the properties developed in Step 2, which, as demonstrated in Step 2, are with high probability  $1 - o(n^{-3/2})$ . For those graphs, we know that for  $k \in (\frac{1}{2} \frac{\log n}{\log(r-1-\epsilon)}, \frac{4}{7} \frac{\log n}{\log(r-1-\epsilon)})$ ,  $p_k \geq (r-1-\epsilon)|Q_k| \geq (r-1-\epsilon)^k \geq n^{1/2}$  and  $p_k \leq r(r-1)^k < n^{2/3-\epsilon}$ . Moreover,  $\sum_{j=1}^k |Q_j| \leq (r-1)|Q_k| = o(n)$  and therefore at the time when  $Q_k$  is generated, there are still  $q_k = n - o(n) = \Theta(n)$  nodes that have not been connected.

Recall that we define the edge configuration between  $Q_k$  and  $Q_{k+1}$  is a list  $C_k = (a_{1,k}, a_{2,k}, \dots)$  where  $a_{i,k}$  means the number of nodes in  $Q_{k+1}$  of which each has exactly  $i$  edges from  $Q_k$ . According to the definition of  $C_k$ , it satisfies

$$\sum_{i=1}^r i \times a_{i,k} = p_k \tag{7}$$

Note that if DE-GNN cannot distinguish  $(u, A^{(1)})$  and  $(u, A^{(2)})$ , then  $A^{(1)}$  and  $A^{(2)}$  must share the same edge configuration between  $Q_k$  and  $Q_{k+1}$ . Otherwise, after one iteration, the intermediate

representation of nodes in  $Q_{k+1}$  are different between  $A^{(1)}$  and  $A^{(2)}$ . Such difference will be propagated to  $u$  later. To bound the probability that  $A^{(1)}$  and  $A^{(2)}$  must share the same edge configuration between  $Q_k$  and  $Q_{k+1}$ , for simplicity, we consider the probability of  $C_k$  given the number of edges between  $Q_k$  and  $Q_{k+1}$ , i.e.,  $p_k$  and the number remaining nodes, i.e.,  $q_k = [n] \setminus \cup_{i=1}^k Q_i = \Theta(n)$ . We are to derive an upper bound of  $\mathbb{P}(C_k)$  based on the configuration model in the following lemma.

**Lemma C.3.** Suppose  $p_k \in [n^{1/2}, n^{2/3-\epsilon}]$  and  $q_k = \Theta(n)$ . Consider the configuration model to generate edges: there are  $p_k$  edges that correspond to two items that are one in  $\cup_{v \in Q_k} W_v$  and one among the rest  $q_k r$  items. Then, for any possible edge configuration  $C_k$  obtained based on this generating procedure,  $\mathbb{P}(C_k) \leq c_5 \frac{q_k^{1/2}}{p_k}$  for some constant  $c_5$ .

*Proof.* First, for the configuration  $C_k = (a_{1,k}, a_{2,k}, \dots)$ , we claim that the most probable  $C_k$  is achieved when  $a_{i,k} = 0$  for  $i \geq 3$ . We prove this statement via the adjustment method: We fix the value of  $a_{1,k} + ia_{i,k}$  and all the other  $a_{i',k}$ 's. We compare the probability of  $(a_{1,k} = x, a_{i,k} = y)$  and that of  $(a_{1,k} = x + i, a_{i,k} = y - 1)$ . Because  $x, y \leq p_k = O(n^{2/3-\epsilon})$ , for some constant  $c_6$ , we have

$$\frac{\mathbb{P}(a_{1,k} = x, a_{i,k} = y)}{\mathbb{P}(a_{1,k} = x + i, a_{i,k} = y - 1)} = \frac{\binom{q_k}{x} r^x \binom{q_k - x}{y} \binom{r}{i}^y}{\binom{q_k}{x+i} r^{x+i} \binom{q_k - x - i}{y-1} \binom{r}{i}^{y-1}} \leq c_6 \frac{x^i}{q_k^{i-1}} \leq c_6 \frac{x^3}{q_k^2} < 1.$$

Therefore, we only need to consider the case when  $a_{1,k}, a_{2,k} > 0$  so  $a_{1,k} + 2a_{2,k} = p_k$ . Define a function  $g(x)$  to denote the probability of the edge configuration  $(a_{1,k} = p_k - 2y, a_{2,k} = y)$ . We compare  $g(y)$  and  $g(y + 1)$

$$\frac{g(y)}{g(y + 1)} = \frac{\binom{q_k}{p_k - 2y} r^{p_k - 2y} \binom{q_k - p_k + 2y}{y} \binom{r}{2}^y}{\binom{q_k}{p_k - 2y - 2} r^{p_k - 2y - 2} \binom{q_k - p_k + 2y + 2}{y+1} \binom{r}{2}^{y+1}} = \frac{2r}{(r-1)} \frac{(y+1)(q_k - p_k + y + 1)}{(p_k - 2y)(p_k - 2y - 1)}. \quad (8)$$

Consider the choice  $y = y^*$  to make  $g(y^*)/g(y^* + 1) \geq 1$  while  $g(y^* - 1)/g(y^*) \leq 1$  that corresponds to  $g(y^*) = \max_y g(y)$ . Then, we must have  $y^* = o(p_k)$  and otherwise  $g(y^* - 1)/g(y^*) > 1$  because  $q_k = \Theta(n)$ . As  $y^* = o(p_k)$  and  $p_k = o(q_k)$ ,  $y^*$  according to Eq. 8 is about  $\frac{(r-1)p_k^2}{2rq_k}$  by setting Eq. 8 = 1. We define  $y_0 = \frac{(r-1)p_k^2}{2rq_k}$ . Consider  $y_0 + \delta$  where  $\delta = o(y_0)$ . Then, using  $p_k = O(n^{2/3-\epsilon})$  and hence  $\frac{y_0 p_k}{q_k} = o(1) = o(\delta)$ , we have

$$\frac{g(y_0 + \delta)}{g(y_0 + \delta + 1)} = 1 + \frac{\delta}{y_0} + o\left(\frac{\delta}{y_0}\right).$$

Moreover, for  $\delta > 0$

$$\begin{aligned} \frac{g(y_0)}{g(y_0 + \delta)} &= \prod_{j=0}^{\delta-1} \left(1 + \frac{j}{y_0} + o\left(\frac{\delta}{y_0}\right)\right) \leq 1 + \frac{\delta(\delta-1)}{2y_0} + o\left(\frac{\delta(\delta-1)}{2y_0}\right), \\ \frac{g(y_0)}{g(y_0 - \delta)} &= \prod_{j=1}^{\delta-1} \left(1 - \frac{j}{y_0} + o\left(\frac{\delta}{y_0}\right)\right)^{-1} \leq 1 + \frac{\delta(\delta-1)}{2y_0} + o\left(\frac{\delta(\delta-1)}{2y_0}\right). \end{aligned}$$

Choose  $\delta = y_0^{1/2}$ , then

$$g(y_0 + y_0^{1/2}), g(y_0 - y_0^{1/2}) \geq \left(\frac{2}{3} + o(1)\right)g(y_0).$$

As  $\sum_{j=-y_0^{1/2}}^{y_0^{1/2}} g(j) \leq 1$  and  $y^*$  should be in  $[y_0 - y_0^{1/2}, y_0 + y_0^{1/2}]$ , we obtain

$$g(y^*) \leq \frac{3}{4y_0^{1/2}} = c_5 \frac{q_k^{1/2}}{p_k},$$

which concludes the proof.  $\square$

Now, we go back to consider any  $\epsilon \frac{\log n}{\log(r-1-\epsilon)}$  many  $k$ 's in  $(\frac{1}{2} \frac{\log n}{\log(r-1-\epsilon)}, \frac{4}{7} \frac{\log n}{\log(r-1-\epsilon)})$ . Based on Lemma C.3, the probability that  $A^{(1)}$  and  $A^{(2)}$  share the same edge configurations between  $Q_k$  and  $Q_{k+1}$  for all these  $k$ 's is bounded by

$$\begin{aligned} \prod_{k=c_7 \frac{\log n}{\log(r-1-\epsilon)}}^{(c_7+\epsilon) \frac{\log n}{\log(r-1-\epsilon)}} \mathbb{P}(C_k) &\leq \prod_{k=c_7 \frac{\log n}{\log(r-1-\epsilon)}}^{(c_7+\epsilon) \frac{\log n}{\log(r-1-\epsilon)}} c_5 \frac{q_k^{1/2}}{p_k} < n^{\epsilon \log c_5} \frac{n^{\frac{\epsilon}{2} \frac{\log n}{\log(r-1-\epsilon)}}}{n^{(c_7+\frac{\epsilon}{2})\epsilon \frac{\log n}{\log(r-1-\epsilon)}}} \\ &< n^{-\frac{\epsilon^2}{3} \frac{\log n}{\log \log n}} = o(n^{-3/2}). \end{aligned}$$

**Step 4:** From step 2, we know that the graphs that do not satisfy  $|Q_k| \geq (r-1-\epsilon)^{k-1}$  and  $p_k \geq (r-1-\epsilon)|Q_k|$  for  $k \in (\frac{\epsilon}{5} \frac{\log n}{\log(r-1)} + 1, (\frac{2}{3} - \epsilon) \frac{\log n}{\log(r-1)})$  are only  $o(n^{-3/2})$  portion of all graphs generated from the configuration model. From step 3, we know that  $A^{(1)}$  and  $A^{(2)}$  that satisfy the properties of step 2, with probability at most  $o(n^{-3/2})$ , their subtrees rooted at node  $u$  are the same even with DE-1's. Even if all these graphs belong to simple regular graphs, as step 1 tells the portion of simple regular graphs among the graphs generated from the configuration model is  $\Omega(n^{-1/2})$ , we arrive at the final conclusion that if  $A^{(1)}$  and  $A^{(2)}$  are sampled from all simple  $r$ -regular graphs, with probability at least  $1 - o(n^{-3/2})/\Omega(n^{-1/2}) = 1 - o(n^{-1})$ , a proper DE-GNN can distinguish  $(u, A^{(1)})$  and  $(u, A^{(2)})$ .  $\square$

## D Discussion on node sets over irregular graphs

Theorem 3.3 focuses on node sets embedded in regular graphs. A natural question therefore arises: how about the power of DEs to distinguish non-isomorphic node sets embedded in irregular graphs that the 1-WL test (also WLGNN) may not distinguish. To answer this question, note that there is some important connection between irregular graphs and regular graphs under the umbrella of the 1-WL test. Actually, the partition of nodes over irregular graphs according to their representations (colors) stably associated by the 1-WL test has *equitable* property [62]. Basically, suppose that the whole node set  $V$  can be partitioned into several parts based on the representations (colors) of nodes,  $V = \cup_{i=1}^c V_i$ , where for an arbitrary  $i$ , nodes in  $V_i$  share the same representations based on the 1-WL test. Then, the induced subgraph of the nodes in  $V_i$  is a regular graph for all  $i$ 's. What's more, for any  $i, j$ , the number of nodes in  $V_j$  that are neighbors of a certain node in  $V_i$  is shared by all the nodes in  $V_i$ , which again shows certain regularity. Theoretically, if we focus on the regular subgraph defined over each  $V_i$ , we may further leverage DE defined over this subgraph to further distinguish the nodes in  $V_i$ . In practice, we do not need to work on each subgraph individually. Mostly, the capability of DE indicated by Theorem 3.3 may still help with breaking such regularity.

## E Proof for The Limitation of DE-1 — Theorem 3.7

We restate Theorem 3.7: Consider any two nodes  $v, u \in V$ . Consider two tuples  $\mathcal{T}_1 = (v, \mathbf{A}^{(1)})$  and  $\mathcal{T}_2 = (u, \mathbf{A}^{(2)})$  with graph topologies  $A^{(1)}$  and  $A^{(2)}$  that correspond to two connected DRGs with a same intersection array. Then, DE-GNN-1 must depend on discriminatory node/edge attributes to distinguish  $\mathcal{T}_1$  and  $\mathcal{T}_2$ .

*Proof.* We are to prove that if  $A^{(1)}$  and  $A^{(2)}$  correspond to two DRGs with a same intersection array, then the subtrees rooted at any nodes are all same even if the nodes are labeled with any DE-1's (see illustration of a subtree rooted at a node in Fig. 3). Because if the subtrees are same, the only possibility to differentiate two nodes is based on discriminatory node/edge attributes embedded in these subtrees when DE-GNN processes them from the bottom to the top.

We recall the definition of the intersection array of a connected DRG as the following.

**Definition E.1.** The intersection array of a connected DRG with diameter  $\Delta$  is an array of integers  $\{b_0, b_1, \dots, b_{\Delta-1}; c_1, c_2, \dots, c_{\Delta}\}$  such that for any node pair  $(u, v) \in V \times V$  that satisfies  $\text{SPD}(v, u) = j$ ,  $b_j$  is the number of nodes  $w$  that are neighbors of  $v$  and satisfy  $\text{SPD}(w, u) = j+1$ , and  $c_j$  is the number of nodes  $w$  that are neighbors of  $v$  and satisfy  $\text{SPD}(w, u) = j-1$ .

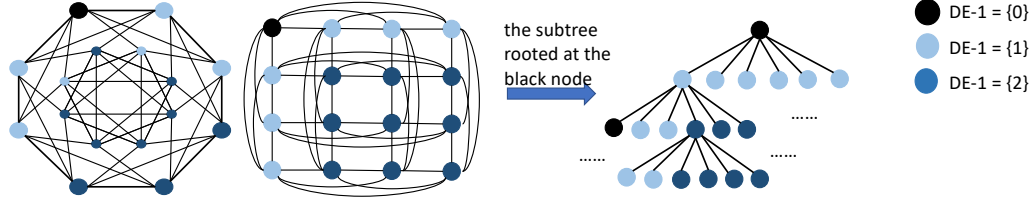


Figure 4: The subtrees rooted at two nodes in the Shrikhande graph and the  $4 \times 4$  rook's graph respectively: In the left two graphs, the black nodes are the target nodes who structural representations are to be learnt. Different colors of the nodes correspond to different DE-1 (SPDs) with respect to the target nodes. For both the Shrikhande graph and the  $4 \times 4$  rook's graph, the subtrees rooted at the black nodes share the same spreading colors as shown in the right. As these two graphs are DRGs with the same intersection array, the configuration of colors (DE-1) of children only depends on the color (DE-1) of their father node.

The definition of DRG implies the following lemma.

**Lemma E.2.** Suppose each node is associated with SPD as DE-1. Consider a graph with the intersection array  $\{b_0, b_1, \dots, b_{\Delta-1}; c_1, c_2, \dots, c_{\Delta}\}$ . For an arbitrary node  $u$ , any node  $v$  in the subtree rooted at  $u$ , with  $\text{SPD}(v, u) = j$ , has children including  $b_j$  nodes  $w$  with DE-1 satisfying  $\text{SPD}(w, u) = j + 1$ ,  $c_j$  nodes with DE-1 satisfying  $\text{SPD} = j - 1$ , and  $b_0 - b_j - c_j$  nodes  $w$  with DE-1 satisfying  $\text{SPD}(w, u) = j + 1$ .

*Proof.* A DRG with the intersection array  $\{b_0, b_1, \dots, b_{\Delta-1}; c_1, c_2, \dots, c_{\Delta}\}$  is a  $b_0$ -regular graph and thus  $v$  has  $b_0$  neighbors. All the neighbors of  $v$  become children of  $v$  in the subtree. As  $\text{SPD}(v, u) = j$ , we know that SPDs from the neighbors of  $v$  to  $u$  are in  $\{j - 1, j, j + 1\}$ . According to the definition of intersection array, we know the numbers of neighbors of  $v$  with different SPDs,  $j - 1, j, j + 1$ , exactly are  $c_j, b_0 - c_j - b_j, b_j$  respectively.  $\square$

We start from any node  $u$  and construct the subtree rooted at  $u$ . By using the Lemma E.2, it is obvious that subtrees for any nodes from DRGs with a same intersection array are all the same even if all the nodes use SPD as DE-1. An illustrative example of this result is shown in Fig. 4.

The next step is to generalize SPD as DE-1 to the list of landing probability as DE-1. Actually, the following lemma indicates that in DRGs, SPD and the list of landing probability are bijective.

**Lemma E.3.** In any connected DRG with the same intersection array, the number of walks of given length between nodes depends only on the SPD between these vertices.

*Proof.* Given any node  $u \in V$  over the DRG with the intersection array  $\mathcal{L} = \{b_0, b_1, \dots, b_{\Delta-1}; c_1, c_2, \dots, c_{\Delta}\}$ , denote the number of walks of length  $l$  from  $u$  to another node  $v$  is  $f(v, l)$ . We need to prove that  $f(v, l)$  can be written as a function  $g(\text{SPD}(u, v), l, \mathcal{L})$ , which is independent from the node identities. We prove this result via induction over the length of walks denoted by  $l$ . The statement is trivial for  $l = 1$ . Suppose the statement is true for  $l = l_0 - 1$ , we consider the case  $l = l_0$ .

Because of the definition of walks, there is recurrence relation

$$\begin{aligned} f(u, v, l_0) &= \sum_{w \in \mathcal{N}_v} f(u, w, l_0 - 1) \\ &= \sum_{w: \text{SPD}(w, u) = \text{SPD}(v, u) - 1} f(u, w, l_0 - 1) + \sum_{w: \text{SPD}(w, u) = \text{SPD}(v, u) + 1} f(u, w, l_0 - 1) \end{aligned}$$

Because the assumption of induction, we have

$$f(u, v, l_0) = c_{\text{SPD}(v, u)} g(\text{SPD}(v, u) - 1, l_0 - 1, \mathcal{L}) + b_{\text{SPD}(v, u)} g(\text{SPD}(v, u) + 1, l_0 - 1, \mathcal{L})$$

which only depends on  $\text{SPD}(v, u), l_0$  and  $\mathcal{L}$  and thus can be written as  $g(\text{SPD}(u, v), l_0, \mathcal{L})$ .  $\square$

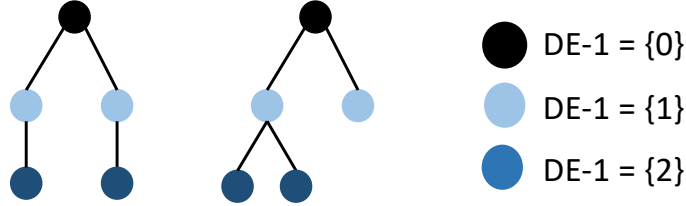


Figure 5:  $K$ -hop aggregation does not necessarily improve the discriminatory power. Consider using DE-GNN-1 and DEA-GNN-1-2-hop to learn the structural representation of the nodes colored by black. We choose SPD as DE-1. Both models require at least two layers to distinguish two black nodes. DEA-GNN-1-2-hop cannot decrease the number of layers by a factor of 2.

Actually, Lemma E.3 is closely related to the argument in [63], which claims the same result for walks over one DRG. However, Lemma E.3 extends the argument to any DRGs with the same intersection array. As DRGs are  $b_0$ -regular graphs, there is a bijective mapping between the list of landing probabilities and the list of walks of different length. Therefore, this is a bijective mapping between SPD and the list of landing probabilities, which concludes the proof.  $\square$

## F Proof for DEA-GNN— Corollary 3.8 and Further Discussion

DEA-GNN contains a general aggregation procedure assisted by DE-1 (Eq. (4)). As we discussed in Section 3.3, as DEA-GNN allows to use DEs as extra node features as DE-GNN in the same time, it has at least the same representation power as DE-GNN. Therefore, Theorem 3.3 and Corollary 3.4 are still true for DEA-GNN. So **our first question** is whether DEA-GNN shares the same limitation with DE-GNN with DE-1 for node representation learning over DRGs. Later, we will prove our confirmation to this question.

An interesting case in practice is to set the DE-1 in Eq. (4) as SPD  $\zeta(u|v) = \zeta_{spd}(u|v)$ . For some  $K \geq 1$ , we specify the aggregation as

$$\text{AGG}(\{f_2(h_u^{(l)}, \mathbf{A}_{vu})\}_{u \in \mathcal{N}_v}) \rightarrow \text{AGG}(\{(f_2(h_u^{(l)}, \mathbf{A}_{vu}), \zeta_{spd}(u|v))\}_{\zeta_{spd}(u|v) \leq K}), \quad (9)$$

which means that the aggregation happens among  $K$ -hop neighbors. If the model is to learn the representation of a node subset with size  $p$ , we term this model as DEA-GNN- $p$ - $K$ -hop. **The second question** is to investigate whether DEA-GNN- $p$ - $K$ -hop may decrease the number of layers of DE-GNN required in Theorem 3.3 and Corollary 3.4 by a factor of  $K$ . This result is not trivial. For example, Fig. 5 shows two trees whose root nodes are the target nodes to learn structural representation. Obviously, DE-GNN-1 needs two layers to distinguish these two root nodes. However, DEA-GNN-1-2-hop may not decrease the number of layers by a factor of 2 and thus still needs two layers. This is because the set aggregation in Eq. (9) may decrease the discriminatory power of those features interacted with graph structures. However, next, we can still prove the confirmation to this second question.

**Confirmation to the first question.** We formally restate the conclusion to prove: Consider any two nodes  $w_1, w_2 \in V$ . Consider two tuples  $\mathcal{T}_1 = (w_1, A^{(1)})$  and  $\mathcal{T}_2 = (w_2, A^{(2)})$  with graph topologies  $A^{(1)}$  and  $A^{(2)}$  that correspond to two connected DRGs with a same intersection array. Then, DEA-GNN-1 must depend on discriminatory node/edge attributes to distinguish  $\mathcal{T}_1$  and  $\mathcal{T}_2$ .

*Proof.* Because Lemma E.3 implies that in all DRGs with the same intersection array, there is a bijective mapping between SPD and the list of landing probabilities. Therefore, we only need to focus on the case that uses SPDs as DE-1, which both appears as extra node attributes and features in aggregation (see Eq. (4)). As the statement is about the limitation, we consider the most expressive case, *i.e.*, the aggregation appearing among every pair of nodes.

Recall the tree structure to compute DE-GNN is termed as the subtree rooted at some node. Now, we call the tree structure to compute DEA-GNN for the structural representation of a node  $w$  as the *extended* subtree rooted at  $w$ , which has the same utility as the tree structure for DE-GNN (Fig. 3).

Suppose we are to learn the structural representation of  $w$ . We are going to prove by induction that with arbitrary number of layers, if no discriminatory node/edge attributes are available, any node  $v$  will be associated with an representation vector that only depends on  $\text{SPD}(v|w)$  and the intersection array of the graph. If this is true, then we know that any nodes in DRGs with a same intersection array will have the same representation output by DEA-GNN with an arbitrary number of layers.

Recall that the number of layers is  $L$ . Obviously, when  $L = 0$ , the node representations only depend on  $\text{SPD}(v|w)$  and thus satisfy the statement. Suppose the statement is true when  $L = L_0 - 1$ , consider the case when  $L = L_0$ . Denote the representation of a node  $v$  after  $L_0 - 1$  layers as  $h_v^{(L_0-1)}$ . Then, its representation after  $L_0$  layers is

$$h_v^{(L_0)} = f_1(h_v^{(L_0-1)}, \text{AGG}\{(f_2(h_u^{(L_0-1)}), \zeta_{\text{spd}}(u|v))\}_{u \in V}).$$

Note that we omit the edge attributes  $\mathbf{A}_{uv}$  due to the requirement of the statement. Consider another node  $v'$  who satisfies  $\text{SPD}(v'|w) = \text{SPD}(v|w)$ . For this, we only need to prove that the following two components are the same and thus  $h_v^{(L_0)} = h_{v'}^{(L_0)}$ :

$$h_v^{(L_0-1)} = h_{v'}^{(L_0-1)}, \quad (10)$$

$$\{(h_u^{(L_0-1)}, \zeta_{\text{spd}}(u|v))\}_{u \in V} = \{(h_u^{(L_0-1)}, \zeta_{\text{spd}}(u|v'))\}_{u \in V}. \quad (11)$$

The Eq. (10) is directly due to the assumption of induction. To prove Eq. (11), we first partition all nodes in  $V$  according to  $\zeta_{\text{spd}}(u|v), \zeta_{\text{spd}}(u|v')$  by defining  $S_v(a) = \{u \in V | \zeta_{\text{spd}}(u|v) = a\}$  and  $S_{v'}(a) = \{u \in V | \zeta_{\text{spd}}(u|v') = a\}$ . We further partition these two sets according to  $\zeta_{\text{spd}}(u|w), \zeta_{\text{spd}}(u|w)$  by defining  $S_v(a, b) = \{u \in S_v(a) | \zeta_{\text{spd}}(u|w) = b\}$  and  $S_{v'}(a, b) = \{u \in S_{v'}(a) | \zeta_{\text{spd}}(u|w) = b\}$ . By using the definition of DRG, we know that  $|S_v(a, b)| = |S_{v'}(a, b)|$  and such cardinality only depends on the intersection array  $\mathcal{L}$ . Moreover, using the assumption of induction, all nodes  $u$  in  $S_v(a, b), S_{v'}(a, b)$  share the same representation  $h_u^{(L_0-1)}$ . Combining the fact that  $|S_v(a, b)| = |S_{v'}(a, b)|$  and the fact the nodes in these two sets hold the same representation, we may claim the second Eq. (11) is true, which concludes the proof.  $\square$

**Confirmation to the second question.** We formally restate the conclusion to prove: Given two fixed-sized sets  $S^{(1)}, S^{(2)} \subset V$ ,  $|S^{(1)}| = |S^{(2)}| = p$ . Consider two tuples  $\mathcal{T}^{(1)} = (S^{(1)}, \mathbf{A}^{(1)})$  and  $\mathcal{T}^{(2)} = (S^{(2)}, \mathbf{A}^{(2)})$  in the most difficult setting when features  $\mathbf{A}^{(1)}$  and  $\mathbf{A}^{(2)}$  are only different in graph topologies specified by  $A^{(1)}$  and  $A^{(2)}$  respectively. Suppose  $A^{(1)}$  and  $A^{(2)}$  are uniformly independently sampled from all  $r$ -regular graphs over  $V$  where  $3 \leq r < (2 \log n)^{1/2}$ . Then, for some constant  $\epsilon > 0$  and constant positive integer  $K$ , there exist a proper DEA-GNN- $p$ - $K$ -hop with layers  $L < \lceil (\frac{1}{2} + \epsilon) \frac{\log n}{K \log(r-1)} \rceil$ , using  $\text{DE-}p \zeta(u|S^{(1)}), \zeta(u|S^{(2)})$  for all  $u \in V$  such that with probability  $1 - o(n^{-1})$ , its outputs  $\Gamma(\mathcal{T}^{(1)}) \neq \Gamma(\mathcal{T}^{(2)})$ .

*Proof.* Similar to the proof of Theorem 3.3, we use Lemma C.1 and focus on the case  $S^{(1)} = S^{(2)} = \{w\}$  and the initial node attribute for each node  $u$  are  $\zeta_{\text{spd}}(u|w)$ . Most of the logic of the proof is the same as that of Theorem 3.3. We only need to take care of the step 3 of the proof of Theorem 3.3, as in DEA-GNN-1- $K$ -hop, a node  $v$  with  $\text{SDP}(v|w) = k$  will aggregation representations of nodes  $u$  even with  $\text{SDP}(u|w) \in [k - K, k + K]$ . Therefore, we need to redefine the edge configuration in the step 3 of the proof of Theorem 3.3.

Recall  $Q_k = \{v \in V | \text{SDP}(v|w) = k\}$ . We define the edge configuration between  $Q_k$  and  $\cup_{i \in [k+1, k+K]} Q_i$  as a list  $\bar{C}_k = ((a_{1,k,1}, a_{2,k,1}, \dots), (a_{1,k,2}, a_{2,k,2}, \dots), \dots, (a_{1,k,K}, a_{2,k,K}, \dots))$ , where  $a_{m,k,j}$  is the number of nodes in  $Q_{k+j}$  of which each connects to exactly  $m$  nodes in  $Q_{k+j-1}$ . Note that two different  $\bar{C}_k$ 's will lead to two different representations of  $(w, A^{(1)})$  and  $(w, A^{(2)})$  after layers  $\lceil \frac{k}{K} \rceil + 2$  as DEA-GNN-1- $K$ -hop uses at most 2 layers yield  $\{h_u^{(2)} | \zeta_{\text{spd}}(u|w, A^{(1)}) = k\} \neq \{h_u^{(2)} | \zeta_{\text{spd}}(u|w, A^{(2)}) = k\}$  and uses at most  $\lceil \frac{k}{K} \rceil$  layers to propagate such difference to the target node  $w$ .

Actually, this definition of edge configuration is nothing but a concatenation of the edge configurations  $\{C_i\}_{i \in [k, k+K-1]}$  where  $C_i$  is the edge configuration between  $Q_i$  and  $Q_{i+1}$  as defined in the proof of Theorem 3.3. Then, we use the statement of step 3 on  $C_i$  to characterize the probabilistic property of



$\bar{C}_k$ . For each  $k \in (\frac{1}{2} \frac{\log n}{\log(r-1-\epsilon)}, \frac{4}{7} \frac{\log n}{\log(r-1-\epsilon)})$ , each type of edge configuration  $\bar{C}_k$  appears with only limited probability  $\mathbb{P}(\bar{C}_k) = \prod_{i=k}^{k+K-1} \mathbb{P}(C_i) = \prod_{i=k}^{k+K-1} O(\frac{n^{1/2}}{p_i})$ . Then, we consider  $\epsilon \frac{\log n}{K \log(r-1-\epsilon)}$  many  $k$ 's in  $(\frac{1}{2} \frac{\log n}{\log(r-1-\epsilon)}, \frac{4}{7} \frac{\log n}{\log(r-1-\epsilon)})$  such that these  $k$ 's hold the same integral interval  $K$  and can be denoted as  $k_0, k_0 + K, \dots$ . The probability that  $A^{(1)}$  and  $A^{(2)}$  have all the same edge configurations for these  $k$ 's is bounded by

$$\prod_{k \in \{k_0, k_0+K, \dots\}} \mathbb{P}(\bar{C}_k) = \prod_{i=k_0}^{k_0+\epsilon \frac{\log n}{\log(r-1-\epsilon)}-1} \mathbb{P}(C_i) \sim o(n^{-\frac{3}{2}}).$$

Therefore, we only need to consider edge configurations for  $k \in (\frac{1}{2} \frac{\log n}{\log(r-1-\epsilon)}, (\frac{1}{2} + \epsilon) \frac{\log n}{\log(r-1-\epsilon)})$  to distinguish  $A^{(1)}$  and  $A^{(2)}$ . And within  $\lceil (\frac{1}{2} + \epsilon) \frac{\log n}{K \log(r-1-\epsilon)} \rceil + 2$ , DEA-GNN-1- $K$ -hop yields different representations for  $(w, A^{(1)})$  and  $(w, A^{(2)})$ . Note that the constant 2 may be merged in  $\epsilon$  for simplicity, which concludes the proof.  $\square$

## G Details of the Experiments

### G.1 Datasets

The three air traffic networks for Task 1, Brazil-Airports, Europe-Airports, and USA-Airports were collected by [64] from the government websites throughout the year 2016 and were used to evaluate algorithms to learn structural representations of nodes [5, 6]. Networks are built such that nodes represent airports and there exists an edge between two nodes if there are commercial flights between them. Brazil-Airports is a network with 131 nodes, 1,038 edges and diameter 5; Europe-Airports is a network with 399 nodes, 5,995 edges and also diameter 5; USA-Airports is a network with 1,190 nodes, 13,599 edges and diameter 8. In each dataset, the airports are divided into 4 different levels according to the annual passengers flow distribution by 3 quantiles: 25%, 50%, 75%. The goal is to infer the level of an airport using solely the connectivity pattern of them.

Tasks 2 & 3 were carried out on three other datasets used by SEAL [9] to facilitate comparison study: C.ele, NS and PB. C.ele [65] is a neural network of *C. elegans* with 297 nodes, 2,148 edges and 3241 triangles (closed node triads), and diameter of 5, in which nodes are neurons and edges are neural linkage between them. NS [66] is a network of collaboration relationship between scientists specialized in network science, comprising of 1461 nodes 2742 edges and 3764 triangles. PB [64] is a network of reference relationships between political post web-pages, consisting of 1222 nodes, 16714 edges, and of diameter 8. Following [9, 10], for Task 2 & 3, we remove all links or triangles in testing sets from graph structure during the training phase to avoid label leakage.

### G.2 Baseline Details

We have five baselines based on GNNs and one baseline, struc2vec [5], based on kernels using handcrafted structural features. We first introduce the implementation of struc2vec and then discuss other baselines.

*Struc2vec* is implemented in a 2-phase manner. In the first phase, embeddings for all the nodes are learned by running the n-gram framework over a constructed graph based on structural similarity kernels. We directly use the code provided by the original paper [5]<sup>4</sup>. In the second phase, the embeddings of nodes that in the target node set are concatenated and further fed into an one-layer fully connected neural network to make further inference.

Regarding other GNN-based baselines, *GCN* is implemented according to Equation (9) of [20] with self-loops added. *SAGE* is implemented according to Algorithm 1 of Section 3.1 in [21]. Mean pooling is used as the neighborhood aggregation function. *GIN* is implemented by adapting the code provided by the original paper [16]<sup>5</sup>, where we use the sum-pooling aggregation and multi-linear perception to aggregate neighbors. In all three baselines described above, ReLU nonlinearities are applied to the output of each hidden layer, followed by a Dropout layer. *PGNN* layer is implemented by adapting the code provided by the original paper [10]<sup>6</sup>. *SEAL* is implemented by adapting the code

<sup>4</sup><https://github.com/leoribeiro/struc2vec>

<sup>5</sup><https://github.com/weihua916/powerful-gnns>

<sup>6</sup><https://github.com/JiaxuanYou/P-GNN>

provided by the original paper <sup>7</sup>. As we focus on learning structural representation with inductive capability, all the five GNN-based methods use node degrees as input features if node attributes are not available.

Final readout layers are tailored to suit different tasks. For Task 1 since the task is node classification, the final layer for all baselines is a one-layer neural network followed by a cross entropy loss. Tasks 2 & 3 have slightly more complex readout layers since the target entity for prediction is a node set of size 2 or 3. Note that SEAL is specifically designed for Task 2 and has its own readout that uses SortPooling over all node representations over the ego-networks of node-pairs [23]. we refer the readers to the original paper for details [9]. For all the other baselines, to make a fair comparison, we use the following difference-pooling: Suppose the target node set is  $S$  and the representation of node  $v$  for  $v \in S$  is denoted by  $h_v$ , then we readout the representation of  $S$  as

$$z = \sum_{u,v \in S} |h_v - h_u| \quad (12)$$

where  $|\cdot|$  denotes component-wise absolute value. Note that Tasks 2 & 3 are to predict the existence of a link / triangle. So we use the inner product  $\langle w, z \rangle$  where  $w \in \mathbb{R}^d$  is a trainable final projection vector, and feed this product into the binary cross entropy loss to train the models.

### G.3 DE-GNN Variants Details.

**Minibatch training based on ego-network extraction.** To understand our detailed framework it is helpful to first discuss the minibatch training of GCN, although the original GCN is trained in a full-batch manner [20]. To train GCN in minibatches, we first extract, for each target node  $v$  in a given minibatch, an ego-network centering at  $v$  within  $L$ -hop neighbors by doing a depth- $L$  BFS from  $v$ , denoted by  $G_v$ . Here,  $L$  is the number of GCN layers intended to be used in the model. Note that the representation of node  $v$  via using  $L$ -layer GCN over  $G_v$  is the same as that via using  $L$ -layer GCN over the whole graph. If node attributes are not available for GCN or other WLGNNs, we may use the degree of each node as its node attributes.

Our models are implemented by following the above mini-batch training framework. For a target node set  $S$ , we first extract the union of ego-networks centering at any nodes in  $S$  within  $L$ -hop neighbors. We call the union of ego-networks as the ego-network around  $S$ , denoted by  $G_S = \cup_{v \in S} G_v$ . Note that even if  $S$  has multiple nodes,  $G_S$  can be extracted as a whole by running BFS. All the edges of  $G_S$  between nodes that are both in  $S$  will be further removed, which is denoted by  $G'_S$ . For  $G'_S$ , we associate each node  $u$  in these ego-networks with the DE  $\zeta(u|S)$  as extra node attributes. Specifically, we use a simple aggregation for Eq. (1):

$$\zeta(u|S) = \frac{1}{|S|} \sum_{v \in S} \zeta(u|v) \quad (13)$$

Next, we detail the different versions of  $\zeta(u|v)$  used by different variants of DE-GNN.

**DE-GNN-SPD.** This variant sets  $\zeta(u|v)$ ,  $v \in S$  and  $u \in G_S$  as a one-hot vector of the truncated shortest-path-distance between  $u$  and  $v$ . That is,

$$\zeta_{spd}(u|v) = \text{one hot}(\min(\text{SPD}(u, v), d_{max})), \quad (14)$$

where  $d_{max}$  is the maximum distance to be encoded. As a result, the  $\zeta_{spd}(u|v)$  is a vector of length  $d_{max} + 1$ . The pairwise SPDs can either be pre-computed in preprocessing stage, or be computed by traversing the extracted ego-networks on the fly. The  $d_{max} (\leq L)$  helps prevent overfitting the noise in an overly large neighborhood.

**Compare DE-GNN-SPD with SEAL.** DE-GNN-SPD, when used for link prediction, is similar to SEAL [9] in a sense that we both encode the distance between any node and the two nodes in the target node-pairs. However, we are fundamentally different from SEAL as SEAL uses graph-level readout of all nodes in the ego-networks. SEAL also has no discussion on the expressive power of distance encoding. Their intention of node labeling as they reported is just to let the model know which node-pair in the extracted ego-networks is the target node-pair. Moreover, the specific DE  $\zeta(u|S)$  for a node-pair  $S = \{v_1, v_2\}$  used in SEAL is a one-hot encoding of the value

<sup>7</sup><https://github.com/muhanzhang/SEAL>

$1 + \min(\text{SPD}(u, v_1), \text{SPD}(u, v_2)) + (d/2)[(d/2) + (d\%2) - 1]$ , where  $d = \text{SPD}(u, v_1) + \text{SPD}(u, v_2)$ . The dimension of this DE is  $O(d_{\max}^2)$  which is higher than our DE (Eq. (14)) used in DE-GNN-SPD, which may result in model overfitting for large  $d_{\max}$

**DE-GNN-LP.** This variant sets  $\zeta(u|v)$ ,  $v \in S$  and  $u \in G_S$  as landing probabilities of random walks (of different lengths) from node  $v$  to node  $u$ :

$$\zeta_{lp}(u|v) = ((W^{(0)})_{vu}, (W^{(1)})_{vu}, (W^{(2)})_{vu}, \dots, (W^{(d_{\text{rw}})})_{vu}) \quad (15)$$

where  $W^{(k)} = (AD^{-1})^k$  is the  $k$ -step random walk matrix,  $d_{\text{rw}}$  is the max step number of random walks. Notice that in principle,  $\zeta_{lp}(u|v)$  encodes the distance information at a finer granularity than  $\zeta_{spd}(u|S)$ . This is because  $\text{SPD}(u, v)$  can be inferred from  $\zeta_{lp}(u|v)$  as the index of the first non-zero random walk feature. However, in practice we observe that such encoding does not always bring a significant performance gain.

For both DE-GNN-SPD and DE-GNN-LP, we use the same 1-hop neighborhood aggregation as GCN. DE can also be used to control the message passing as shown in Eq. (4). We further discuss two variants by incorporating DE-GNN-SPD and Eq. (4).

**DEA-GNN-SPD.** DEA-GNN-SPD is built on top of DE-GNN-SPD, using the same extra node features  $\zeta_{spd}(u|S)$ , but allows for multi-hop aggregation by specifying Eq. (4) as

$$\text{AGG}(\{f_2(h_u^{(l)}, \mathbf{A}_{vu})\}_{u \in \mathcal{N}_v}) \rightarrow \text{AGG}(\{(f_2(h_u^{(l)}, \mathbf{A}_{vu}), \zeta_{spd}(u|v))\}_{\zeta_{spd}(u|v) \leq K}).$$

In experiments, we choose  $K = 2, 3$ , which means that each node aggregates representations of other nodes that are not only its direct neighbors but also its (exclusive) 2-hop and even 3-hop neighbors. As we do not have edge attributes in our data, we omit  $\mathbf{A}_{vu}$ . Our implementation of the aggregation for the layer  $l$  follows

$$h_v^{(l+1)} = \sum_{k=1}^K \text{Relu} \left( \frac{1}{|S_{v,k}| + 1} \left( h_v^{(l)} + \sum_{u \in S_{v,k}} h_u^{(l)} \Theta^{(lk)} \right) \right), \quad S_{v,k} = \{u | \zeta_{spd}(u|v) = k\}.$$

where  $\Theta^{(lk)}$  is a trainable weight matrix and for each  $k$ , we aggregate  $k$ -hop neighbors via a GCN layer with a self-loop. Note that when implementing DEA-GNN-SPD, we need to extract the ego-network of nodes within  $LK$ -hops, if DEA-GNN-SPD has  $L$  layers.

**DEA-GNN-PR.** DEA-GNN-SPD is also built on top of DE-GNN-SPD, but the propagation is by specifying Eq. (4) as

$$\text{AGG}(\{f_2(h_u^{(l)}, \mathbf{A}_{vu})\}_{u \in \mathcal{N}_v}) \rightarrow \text{AGG}(\{(f_2(h_u^{(l)}, \mathbf{A}_{vu}), \zeta_{gpr}(u|v))\}_{v \in V}).$$

As the aggregation is over the whole node set, this model does not extract the ego-networks but uses the entire graphs. For the layer  $l$ , we further specify the above aggregation by using

$$h_v^{(l+1)} = \text{Relu} \left( \sum_{u \in V} \zeta_{gpr}(u|v) h_u^{(l)} \Theta^{(l)} \right)$$

where  $\Theta^{(l)}$  is a trainable weight matrix and  $\zeta_{gpr}(u|v)$  is a specific form of  $\zeta_{gpr}(u|v)$  based on Personalized PageRank scores [67], i.e.,

$$\zeta_{gpr}(u|v) = \left[ \sum_{k=0}^{\infty} (0.9W)^k \right]_{uv} = [(I - 0.9W)^{-1}]_{uv}.$$

Note that the above 0.9 is a hyper-parameter. As we are just willing to show the use case, 0.9 is set as a heuristic and is not obtained via parameter tuning. Other values may yield better performance. Other types of PageRank scores may be used, *e.g.*, heat-kernel PageRank scores [68], time-dependent PageRank scores [69].

We compare DEA-GNN-PR with all other methods, which yields the following Table 2. DEA-GNN-PR performs worse than DEA-GNN-SPD while it still works much better than WLGNs in link and triangle predictions. Comparing these observations with the statements on GDC [32], we argue that missing DEs as node attributes is the key that limits the performance of link prediction via GDC.

Method	Data	Nodes (Task 1): Average Accuracy			Node-pairs (Task 2): AUC			Node-triads (Task 3): AUC	
		Bra.-Airports	Eur.-Airports	USA-Airports	C.elegans	NS	PB	C.elegans	NS
GCN [20]		64.55±4.18	54.83±2.69	56.58±1.11	74.03±0.99	74.21±1.72	89.78±0.99	80.94±0.51	81.72±1.50
SAGE [21]		70.65±5.33	56.29±3.21	50.85±2.83	73.91±0.32	79.96±1.40	90.23±0.74	84.72±0.40	84.06±1.14
GIN [16]		71.89±3.60†	57.05±4.08	58.87±2.12	75.58±0.59	87.75±0.56	91.11±0.52	86.42±1.12†	94.59±0.66†
Struc2vec [5]		70.88±4.26	57.94±4.01†	61.92±2.61†	72.11±0.31	82.76±0.59	90.47±0.60	77.72±0.58	81.93±0.61
PGNN [10]		N/A	N/A	N/A	78.20±0.33	94.88±0.77	89.72±0.32	86.36±0.74	79.36±1.49
SEAL [9]		N/A	N/A	N/A	88.26±0.56†	98.55±0.32†	94.18±0.57†	N/A	N/A
DE-GNN-SPD		<b>73.28±2.47</b>	56.98±2.79	<b>63.10±0.68*</b>	<b>89.37±0.17*</b>	<b>99.09±0.79</b>	<b>94.95±0.37*</b>	<b>92.17±0.72*</b>	<b>99.65±0.40*</b>
DE-GNN-LP		<b>75.10±3.80*</b>	58.41±3.20*	<b>64.16±1.70*</b>	86.27±0.33	98.01±0.55	91.45±0.41	86.24±0.18	<b>99.31±0.12*</b>
DEA-GNN-SPD		<b>75.37±3.25*</b>	57.99±2.39*	<b>63.28±1.59</b>	<b>90.05±0.26*</b>	<b>99.43±0.63*</b>	<b>94.49±0.24*</b>	<b>93.35±0.65*</b>	<b>99.84±0.14*</b>
DEA-GNN-PR		<b>73.26±4.08</b>	51.41±2.39	50.34±1.50	83.07±0.77	<b>99.46±0.37*</b>	92.68±0.57	83.15±1.11	<b>99.86±0.03*</b>

Table 2: Model performance (including DEA-GNN-PR) in Average Accuracy and Area Under the ROC Curve (AUC) (mean in percentage  $\pm$  95% confidence level). † highlights the best baselines. \*, **bold font**, **bold font\*** respectively highlights the case where our proposed model’s performance: exceeds the best baseline on average, exceeds by 70% confidence, exceeds by 95% confidence.

#### G.4 Model performance without validation datasets

We have confirmed with the original authors of Struc2vec [5] and SEAL [9] that the performance of these two baselines reported in their papers do not use validation set. The performance therein is the best testing results ever achieved when models are being trained till convergence. We think that it is necessary to include validation datasets to achieve fair comparison and therefore reported the results with validation datasets in the main text. We put the results without validation datasets here. Under both experimental settings, we draw similar conclusions for our models in comparison with the baselines.

Method	Data	Nodes (Task 1): Average Accuracy			Node-pairs (Task 2): AUC			Node-triads (Task 3): AUC	
		Bra.-airports	Eur.-airports	USA-airports	C.elegans	NS	PB	C.elegans	NS
GCN [20]		82.01±3.09	55.56±2.90	60.08±4.20	78.10±1.24	81.92±0.90	90.04±1.23	83.15±0.50	88.56±3.54
GraphSAGE [21]		81.48±6.26	63.41±5.87	54.62±4.29	77.20±1.26	86.16±2.95	90.73±1.38	86.84±1.13	87.98±4.58
GIN [16]		85.19±5.85†	65.47±2.99†	63.45±5.29	79.79±1.42	89.84±3.62	91.47±0.58	89.37±0.44†	95.48±0.64†
stuc2vec [5]		84.08±9.30	65.30±5.60	68.09±2.50†	74.17±0.37	87.43±2.48	90.96±0.26	83.66±3.26	87.53±3.12
PGNN [10]		N/A	N/A	N/A	80.76±0.98	94.99±1.44	90.21±0.78	87.18±0.52	83.39±1.74
SEAL [9]		N/A	N/A	N/A	90.30±1.35†	98.85±0.47†	94.72±0.46†	N/A	N/A
DE-GNN-SPD		<b>87.78±3.95</b>	65.62±3.27*	<b>70.95±0.80*</b>	90.67±0.91*	<b>99.50±0.87</b>	<b>95.21±0.53</b>	<b>92.54±0.63*</b>	<b>99.91±0.13*</b>
DE-GNN-LP		<b>88.36±4.62</b>	66.25±2.07*	69.33±1.33*	88.48±0.46	98.24±0.67	92.43±1.29	87.65±0.71	<b>99.74±0.21*</b>
DEA-GNN-SPD		<b>88.43±3.24*</b>	<b>67.14±1.57*</b>	<b>71.07±0.63*</b>	<b>91.09±0.96*</b>	<b>99.52±0.88</b>	<b>95.15±0.24*</b>	<b>93.28±0.51*</b>	<b>99.94±0.06*</b>
DEA-GNN-PR		<b>87.30±2.52</b>	64.38±1.78	62.46±1.18	84.23±0.55	<b>99.58±0.46*</b>	92.94±0.76	84.61±0.58	<b>99.95±0.06*</b>

Table 3: Model performance without validation set (including DEA-GNN-PR) in Average Accuracy and Area Under the ROC Curve (AUC) (mean in percentage  $\pm$  95% confidence level). † highlights the best baselines. \*, **bold font**, **bold font\*** respectively highlights the case where our proposed model’s performance: exceeds the best baseline on average, exceeds by 70% confidence, exceeds by 95% confidence.

We report additional results of our model on Task 2 and 3 versus other baselines, measured by average accuracy without validation set in Table 4. Similar observations as reported in the main text can be drawn from both Table 3 and 4: the strongest baselines are given by SEAL [9] and GIN [16], while our DEA-GNN variants further significantly outperform those baselines on all tasks.

Method	Data	Node-pairs (Task 2) Average Accuracy			Node-triads (Task 3) Average Accuracy	
		C.elegans	NS	PB	C.elegans	NS
GCN [20]		65.01±0.45	74.69±1.39	78.35±1.69	69.47±2.69	82.16±1.27
SAGE [21]		67.96±0.90	78.29±2.40	83.53±1.41	76.67±0.72	88.45±0.65
GIN [16]		69.45±1.24	82.17±1.07	83.01±0.83	77.60±0.85†	92.67±1.44†
Struc2vec [5]		68.41±2.57	80.64±0.92	74.43±3.48	71.06±3.86	83.42±1.67
PGNN [10]		71.40±1.68	91.04±0.85	86.44±1.14	76.34±0.23	80.67±0.49
SEAL [9]		83.90±0.97†	97.89±0.40†	88.92±0.95†	N/A	N/A
DE-GNN-SPD		83.78±0.88	<b>99.54±0.28*</b>	<b>89.82±0.82*</b>	<b>90.82±0.62*</b>	<b>100.0±0.00*</b>
DE-GNN-LP		70.96±1.27	96.96±0.57	84.37±1.03	<b>82.90±0.69*</b>	<b>100.0±0.00*</b>
DEA-GNN-SPD		<b>84.81±0.69*</b>	<b>99.77±0.24*</b>	<b>89.83±0.98*</b>	<b>91.24±0.22*</b>	<b>100.0±0.00*</b>
DEA-GNN-PR		76.95±0.85	<b>99.88±0.19*</b>	85.82±0.87	<b>83.18±0.71*</b>	<b>100.0±0.00*</b>

Table 4: Model performance without validation set (including DEA-GNN-PR) in Average Accuracy (mean in percentage  $\pm$  95% confidence level). † highlights the best baselines. \*, **bold font**, **bold font\*** respectively highlights the case where our proposed model’s performance: exceeds the best baseline on average, exceeds by 70% confidence, exceeds by 95% confidence.

#### G.5 Hyperparameters Tuning.

Table 5 lists the most important hyperparameters’ at a glance, which applies to both the baselines and our proposed models. Grid search is used to find the best hyperparameters combination. The models

are sufficiently trained till the cross entropy loss converges and we report the best model by running each for 20 times over different random seed. For more details please refer to the code attached.

Hyperparameters	Value / Range	Notes
batch size	64, 128	
learning rate	1e-4	
optimizer	SGD	stochastic gradient descent
conv. layers	1, 2, 3	struc2vec does not follow this setting
conv. hidden dim.	20, 50, 80, 100	struc2vec does not follow this setting
dropout	0, 0.2	
$d_{rw}$	3, 4	#steps of random walk, valid only for DE-GNN-RW
$d_{spd}$	3, 4	maximum shortest path distance for DE-GNN-SPD variants
prop_depth	1, 2, 3	the number of hops of message in one layer, only valid for DEA-GNN-SPD, 1 for all the others

Table 5: List of hyperparameters and their value / range.

## H A Brief Introduction of Higher-Order Weisfeiler-Lehman Tests

We mentioned higher-order WL tests in our context, especially on the 2-WL test. There are different definitions of the  $k$ -WL test ( $k \geq 2$ ), while in this work we followed the definition in [39]. Note that the  $k$ -WL test here also corresponds to the  $k$ -WL' test in [40] and the  $k$ -FWL test in [27], and are equivalent to the  $k + 1$ -WL tests in [26, 27].

The  $k$ -WL test ( $k \geq 2$ ) follows the following coloring procedure:

1. For each  $k$ -tuple of node set  $V_i = (v_{i_1}, v_{i_2}, \dots, v_{i_k}) \in V^k$ ,  $i \in [n^k]$ , we initialize  $V_i$  with a color denoted by  $C_i^{(0)}$ . These colors satisfies that for two  $k$ -tuples,  $V_i$  and  $V_j$ ,  $C_i^{(0)}$  and  $C_j^{(0)}$  are the same if and only if for  $a, b \in [k]$  (1)  $v_{i_a} = v_{i_b} \Leftrightarrow v_{j_a} = v_{j_b}$  and (2)  $(v_{i_a}, v_{i_b}) \in E \Leftrightarrow (v_{i'_a}, v_{i'_b}) \in E$ .
2. For each  $k$ -tuple  $V_i$  and  $u \in V$ , define  $N(V_i; u)$  as a sequence of  $k$ -tuples such that  $N(V_i; u) = ((u, v_{i_2}, \dots, v_{i_k}), (v_{i_1}, u, \dots, v_{i_k}), (v_{i_1}, v_{i_2}, \dots, u))$ . Then, the color of  $V_i$  can be updated via the following mapping:

$$C_i^{(l+1)} \leftarrow g(C_i^{(l)}, \{(C_j^{(l)} | V_j \in N(V_i; u))\}_{u \in V}),$$

where  $g(\cdot)$  is injective coloring.

3. For each step  $l$ ,  $\{C_i^{(l)}\}_{i \in [n^k]}$  is a coloring configuration of the graph  $G$ , which is essentially a multi-set. If two graphs have different coloring configurations, these two graphs are determined to be non-isomorphic, while the inverse is not true.

Note that the step 2 essentially requires to aggregate colors of  $nk$ -tuples and thus even when  $k = 2$ , the 2-WL test may not leverage the sparsity of graph structure to keep good scalability. Ring-GNN [30] and PPGN [27] essentially try to achieve the expressive power of the 2-WL test. They are not scalable to process large graphs and they were only evaluated for entire-graph-level tasks such as graph classification and graph regression [27, 30].