Zero-Shot Generalization of GNNs over Distinct Attribute Domains

Yangyi Shen [1] **Jincheng Zhou** [2] **Beatrice Bevilacqua** [2] **Joshua Robinson** [1] **Charilaos Kanatsoulis** [1]
Jure Leskovec [1] **Bruno Ribeiro** [2]

Abstract

Traditional Graph Neural Networks (GNNs) cannot generalize to new graphs with node attributes different from the training ones, making zero-shot generalization across different node attribute domains an open challenge in graph machine learning. In this paper, we propose STAGE, which encodes *statistical dependencies* between attributes rather than individual attribute values, which may differ in test graphs. By assuming these dependencies remain invariant under changes in node attributes, STAGE achieves provable generalization guarantees for a family of domain shifts. Empirically, STAGE demonstrates strong zero-shot performance on medium-sized datasets: when trained on multiple graph datasets with different attribute spaces (varying in types and number) and evaluated on graphs with entirely new attributes, STAGE achieves a relative improvement in Hits@1 between 40% to 103% in link prediction and a 10% improvement in node classification compared to state-of-the-art baselines.

1. Introduction

Zero-shot generalization refers to the ability of the model to handle unseen test data without additional training or adaptation (Larochelle et al., 2008; Xian et al., 2017; Wang et al., 2022). An essential prerequisite for zero-shot generalization is a unified input space where models can learn and transfer prediction patterns across domains. While this challenge has been addressed in areas like natural language through tokenization techniques that represent any text through a fixed vocabulary (Samuel & Øvrelid, 2023), graphs present unique challenges in achieving such unified input space.

Attributes in graphs can vary significantly across domains.

[1]Department of Computer Science, Stanford University, Stanford, USA [2]Department of Computer Science, Purdue University, West Lafayette, USA. Correspondence to: Yangyi Shen <pyyshen@stanford.edu>.

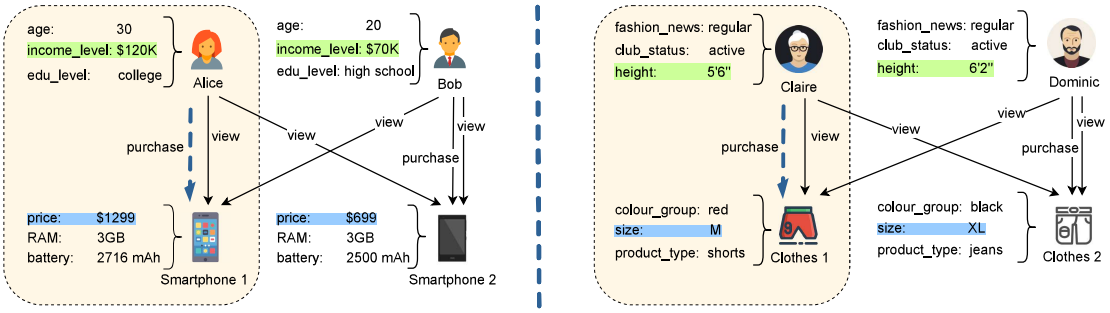
Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

out-of-distribution

Node attributes in test graphs can differ from those in training graphs in four key ways: (1) their types (e.g., continuous vs. categorical variables); (2) their names (e.g., *RAM* specifications in ecommerce graphs and clothing *size* in retail graphs, as illustrated in Figure 1); (3) their semantics, where attributes with the same name can have different meanings across domains – for instance, the meaning of *size* differs substantially between electronics and clothing domains; (4) their cardinality, as graphs may contain varying numbers of node attributes. *These challenges make it difficult to define a unified input space that enables zero-shot generalization to unseen attributed graphs.*

For these reasons, training graph models that can zero-shot generalize to new graphs with unseen attribute domains remains an open challenge. Recent approaches address this problem using various strategies. One approach is to ignore node attributes to focus solely on graph topology, but this strategy may be leaving valuable node attribute information unutilized. Another line of work seeks to unify input spaces by converting graphs and attributes into text representations, which are then processed by pretrained text encoders (Chen et al., 2024a; Huang et al., 2023; Liu et al., 2024; Zhang et al., 2023). While promising, these approaches may struggle with numerical attributes (Collins et al., 2024; Gruver et al., 2024; Schwartz et al., 2024). Recently, Zhao et al. (2024b) proposed an analytical approach for making predictions on new graphs with potentially new attributes. However, this approach sidesteps the fundamental challenge of creating a unified input space.

In this paper, we introduce STAGE (**S**tatistical **T**ransfer for **A**ttributed **G**raph **E**mbeddings), which transforms node attributes from their "absolute" natural space into a *relative* space that captures statistical dependencies between attributes. For instance, as illustrated in Figure 1, these dependencies manifest themselves as correlations driving purchases across domains, which remain invariant even when the purchased items and their attributes change. In practice, STAGE represents such statistical dependencies through a two-step process that transforms node attributes into fixed-dimensional edge embeddings, achieving a unified input space alongside provable invariance to changes in attribute values (including their types, names and semantics), as well as to permutations of attribute order and permuta-



(a) Train (b) Test

Figure 1: The task of zero-shot generalization to attributed graphs with unseen attributes. Attributes in test are different than those in train in types and semantics, but attributes associated with an edge are highly correlated in both train and test (e.g. income level is positively correlated to phone price in (a) and height is positively correlated to size in (b)). Our STAGE learns these *statistical dependencies* among attributes to perform zero-shot transfer across distinct attribute domains.

tions of node identities. Specifically, STAGE first constructs a weighted *STAGE-edge-graph* for each edge in the input graph, where the nodes represent attributes of the edge endpoints and the edge weights capture dependencies between the attributes. Then, STAGE uses an additional shallow GNN to generate embeddings for each *STAGE-edge-graph*. Finally, STAGE applies the original GNN to a modified input graph, which contains only the newly generated edge embeddings but not the node attributes.

The complexity of STAGE is linear in the size of the input graph and quadratic in the number of attributes, as it captures pairwise statistical dependencies between attributes over the edges of the graph. This makes STAGE particularly well-suited for small to medium-sized datasets, where it strikes a balance between computational feasibility and strong generalization performance.

We prove that STAGE can learn domain-independent representations for certain types of domain shifts, enabling zero-shot generalization. Experimentally, for link prediction in e-commerce networks spanning six distinct product domains, STAGE achieves up to 103% improvement in Hits@1 compared to the strongest baseline. In node classification tasks on social networks, STAGE achieves approximately 10% better performance than the strongest baseline.

2. STAGE

Let $G = (V, E, \boldsymbol{X})$ an attributed graph, where V is the set of nodes, E the set of edges, and $\boldsymbol{X} = \{\boldsymbol{x}^v\}_{v \in V}$ the set of node attributes $\boldsymbol{x}^v$ for each node $v \in V$. We assume that all $\boldsymbol{x}^v$ belong to some measurable space of dimension $d \geq 1$.

To design a model capable of generalizing to test graphs that may have node attributes living in a different space than $\boldsymbol{X}$, we propose a projection map that transforms the node attributes $(\boldsymbol{x}^u, \boldsymbol{x}^v)$ of the endpoints of an edge $(u, v) \in E$ into a fixed-dimensional pairwise embedding

$$\mathcal{P} : (\boldsymbol{x}^u, \boldsymbol{x}^v) \mapsto \boldsymbol{r}^{uv} \in \mathbb{R}^k, \quad k \geq 1. \tag{1}$$

By using pairwise embeddings, STAGE can *model relationships between attributes belonging to different nodes*. For instance, it can capture the relation between the attributes of the customer node Alice and the attributes of the product node Phone1 in Figure 2(a), such as the correlation between income level and price. We design the mapping $\mathcal{P}$ by building a graph based on the pairwise *pdf* attribute descriptors. Viewing node attributes through their *pdf*s maps potentially non-aligned node attribute spaces into a universal space of densities, enabling consistency across diverse domains. The modeling of probabilities generalizes the learning of rules like "people with higher income level tend to buy expensive phones," to abstract relationships like "high values in X_1 correlate with high values in X_2", enabling knowledge transfer across domains with different attributes.

Concretely, let A and B be a random pair of nodes jointly and uniformly sampled from the edge set, $(\mathrm{A}, \mathrm{B}) \sim \mathrm{Unif}(E)$. Let $\mathrm{x}_i^{\mathrm{A}}$ denote the random variable of the i-th attribute value of random node A, and $\mathrm{x}_j^{\mathrm{B}}$ the j-th attribute value of random node B. Given a specific pair of distinct nodes $u, v \in V$ and specific attribute values x_i^u and x_j^v, we define $p(x_i^u | x_j^v)$ from the conditional probabilities as follows, accounting for mixture of totally ordered (e.g., scalar) and unordered (e.g., categorical) attributes[1]:

- $p(x_i^u | x_j^v) := \mathbb{P}(\mathrm{x}_i^{\mathrm{A}} \leq x_i^u | \mathrm{x}_j^{\mathrm{B}} \leq x_j^v)$, if both attribute i and j are totally ordered.
- $p(x_i^u | x_j^v) := \mathbb{P}(\mathrm{x}_i^{\mathrm{A}} = x_i^u | \mathrm{x}_j^{\mathrm{B}} \leq x_j^v)$, if attribute i is unordered and attribute j is totally ordered.
- $p(x_i^u | x_j^v) := \mathbb{P}(\mathrm{x}_i^{\mathrm{A}} \leq x_i^u | \mathrm{x}_j^{\mathrm{B}} = x_j^v)$, if attribute i is totally ordered and attribute j is unordered.

[1]For brevity we omit the distribution $(\mathrm{A}, \mathrm{B}) \sim \mathrm{Unif}(E)$, writing $\mathbb{P}$ instead of $\mathbb{P}_{(\mathrm{A},\mathrm{B})\sim\mathrm{Unif}(E)}$ from now.

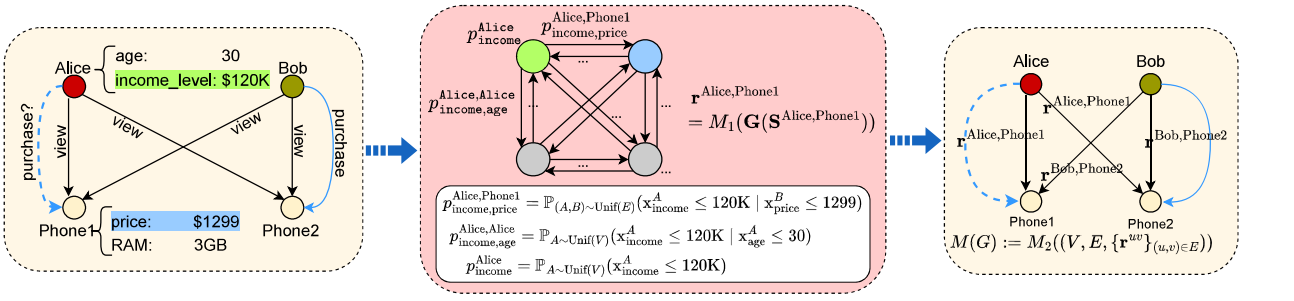


Figure 2: Given an input attributed graph G (a), STAGE builds a *STAGE-edge-graph* (b) for every edge in G. Nodes in a STAGE-edge-graph correspond to attributes of the two edge endpoints, and the node and edge attributes are the empirical marginal and conditional probabilities of attribute values (Equations (2) and (3)). STAGE applies the intra-edge GNN on STAGE-edge-graphs (b) to obtain an edge embedding for each input graph edge, and then applies the inter-edge GNN on the modified graph containing these edge embeddings but not the node attributes (c). Details are provided in Algorithms 1 and 2.

- $p(x_i^u|x_j^v) := \mathbb{P}(\mathrm{x}_i^{\mathrm{A}} = x_i^u | \mathrm{x}_j^{\mathrm{B}} = x_j^v)$, if both attribute i and j are unordered.

If $u = v$, we change the sampling distribution to $(\mathrm{A}) \sim \mathrm{Unif}(V)$ and let $\mathrm{B} = \mathrm{A}$ so that STAGE can also model dependencies between attributes of the same node. If $i = j$, we change the conditional probability to $p(x_i^u) := \mathbb{P}(\mathrm{x}_i^A = x_i^u)$ if attribute i is unordered and $p(x_i^u) := \mathbb{P}(\mathrm{x}_i^A \leq x_i^u)$ if attribute i is totally ordered. This allows STAGE to also model each attribute independently through its *pdf* or *cdf*.

In practice, these probabilities can be empirically estimated from the input data. For the node-pair u, v we define a conditional probability matrix $\boldsymbol{S}^{uv}$, with indices $i, j \in \{1, \ldots, 2d\}$, $i \neq j$, organized such that indices 1 to d correspond to attributes of node u and indices $d+1$ to $2d$ correspond to attributes of node v:

$$\boldsymbol{S}_{ij}^{uv} = \begin{cases} p(x_i^u \mid x_j^u) & \text{if } i \leq d \text{ and } j \leq d, \\ p(x_{i-d}^v \mid x_{j-d}^v) & \text{if } d < i \leq 2d \text{ and } d < j \leq 2d, \\ p(x_i^u \mid x_{j-d}^v) & \text{if } i \leq d \text{ and } d < j \leq 2d, \\ p(x_{i-d}^v \mid x_j^u) & \text{if } d < i \leq 2d \text{ and } j \leq d. \end{cases} \tag{2}$$

and for the diagonal $i = j$ we define,

$$\boldsymbol{S}_{ij}^{uv} = \begin{cases} p(x_i^u) & \text{if } i \leq d, \\ p(x_i^v) & \text{if } i > d, \end{cases} \tag{3}$$

The matrix $\boldsymbol{S}^{uv}$ is the core node-pair data representation STAGE uses. This matrix is used to define a graph structure which we call a STAGE-edge-graph, illustrated in Figure 2(b), which captures, for the pair of nodes u and v, the interactions among all pairs of attributes.

Definition 2.1 (STAGE-edge-graph)**.** Given a pair of nodes $u, v \in V$, a STAGE-edge-graph for (u, v) is a fully connected, weighted, directed graph $\boldsymbol{G}(\boldsymbol{S}^{uv})$ with $2d$ nodes, where node i has a scalar attribute $\boldsymbol{S}_{ii}^{uv}$, and edge (i, j) has a scalar attribute $\boldsymbol{S}_{ij}^{uv}$.

STAGE algorithm. As illustrated in Figures 2(b) and 2(c), STAGE uses a STAGE-edge-graph for each edge in the input graph in a two-stage process to produce attribute-domain-transferable representations. First, STAGE uses a GNN to obtain embeddings for each STAGE-edge-graph. These edge embeddings replace the original node attributes, resulting in a modified graph which is fed into a second GNN to solve the overall task, producing node, link, or graph representation. The two steps of STAGE are as follows:

1. *(Intra-edge)* Each $\boldsymbol{G}(\boldsymbol{S}^{uv})$ is processed with a GNN M_1 to produce edge-level embeddings $\boldsymbol{r}^{uv} = M_1(\boldsymbol{G}(\boldsymbol{S}^{uv}))$.
2. *(Inter-edge)* A second GNN M_2 processes the modified graph $G' = (V, E, \{\boldsymbol{r}^{uv}\}_{(u,v)\in E})$, i.e., the original graph without node attributes, but equipped with the learned edge embeddings to give a final representation $M(G) := M_2(G')$.

The two GNNs M_1 and M_2 are trained end-to-end on the task. Note that M_1 can be any GNN designed to produce whole-graph embeddings and can take single-dimensional edge attributes, whilst M_2 can be any GNN that can take edge embeddings as input.

Integration with language models. While STAGE can incorporate LLM embeddings for textual attributes, our experiments show STAGE-edge-graphs performs better on numerical and categorical data (Section 4). The approaches can be complementary - initialize node embeddings with

LLM embeddings for textual attributes and edge embeddings with STAGE-edge-graphs for non-textual attributes.

Modelling pairwise relations. $\boldsymbol{S}^{uv}$ is only computed for *edges* (u, v), and so can only model pairwise relations between nodes connected by an edge. In some cases, such as bipartite graphs, we find it beneficial to add extra edges between nodes of the same type (see Section 4 for details). In general, higher-order relations could also be modelled similarly, albeit at increased complexity. We leave exploration of higher-order relations to future work.

3. Statistical Underpinnings of STAGE

This section explains how STAGE achieves domain transferability. The central result is to show that STAGE generates representations capable of measuring dependencies among node attributes in graphs. This means that STAGE can ignore "absolute" attribute values, while still generalizing through analogous statistical dependencies of the attributes.

Our first step (Section 3.1) connects measures of statistical dependencies with a novel graph regression task. Then, Section 3.2 shows that our STAGE-edge-graphs (Definition 2.1) can lead to a compact model for this regression, with a variant that is invariant to a class of shifts between train and test attribute domains. The following theoretical results are meant to provide insights and are restricted to domains with a fixed number of attributes to simplify the proofs, extending them to variable size spaces is left as future work. Detailed proofs are provided in Appendix B.

3.1. Statistical Dependence as Graph Regression

We begin by introducing the framework for building what we call *feature hypergraphs*. We will show that feature hypergraphs can sufficiently encapsulate the statistical dependencies between attributes, while only leveraging the relative orders rather than the numerical values of the attribute, enabling it to be invariant to order-preserving transformations (formally defined in Definition B.2) to achieve domain transferability. In the following, we assume one attribute space defined over a totally ordered set (e.g., $\mathbb{R}^d$ for $d \geq 1$, where the total order $\leq_\tau$ is well defined), since the invariances of unordered sets are a special case (as these do not need order-preserving transformations). Before we describe how feature hypergraphs are built, we start with the concept of order statistic, which captures the relative ordering of the attribute values.

Order statistic (David & Nagaraja, 2004). Let $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m$ be a sequence of $m \geq 2$ random variables from some unknown distribution F over a totally ordered set (e.g., a convex set $\mathbb{F} \subseteq \mathbb{R}$). Its *order statistics* are defined as the sorted values $\mathbf{x}_{(1)} \leq \mathbf{x}_{(2)} \leq \cdots \leq \mathbf{x}_{(m)}$, where $\mathbf{x}_{(k)}$ denotes the k-th smallest value in the m samples.

Consider a domain with m entities (e.g., products in an appliance store), where each entity is characterized by d attributes. Specifically, an entity u can be represented by a (row) vector of random attribute variables, $\mathbf{x}^u = [\mathrm{x}_1^u, \mathrm{x}_2^u, \ldots, \mathrm{x}_d^u]$, where x_i^u describes the i-th attribute of entity u that takes on values from the i-th attribute space $\mathbb{F}_i \subseteq \mathbb{R}$. With these variables, we define the (random) matrix $\mathbf{X} \coloneqq [(\mathbf{x}^1)^T, (\mathbf{x}^2)^T, \ldots, (\mathbf{x}^m)^T]^T$ of shape $m \times d$. Alternatively, we can view $\mathbf{X}$ column-wise, where each attribute i corresponds to a (column) random vector $\mathbf{x}_i = [\mathrm{x}_i^1, \mathrm{x}_i^2, \ldots, \mathrm{x}_i^m]^T$. Next, we introduce the order statistic for these attributes: let $\mathbf{x}_{i(k)}$ denote the *k-th order statistic* of $\{\mathrm{x}_i^1, \ldots, \mathrm{x}_i^m\}$. For instance, $\mathbf{x}_{i(1)} = \min\{\mathrm{x}_i^1, \ldots, \mathrm{x}_i^m\}$.

Given an input graph $G = (V, E, \boldsymbol{X})$, we regard it as a sample from some unknown distribution over all attributed graphs with m entities and d attributes, where $\boldsymbol{X}$ is a random variable with $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_d]$. Consider the edges in E as samples of pairs of nodes that give rise to the multiset of attributes of the endpoint nodes, $\mathcal{E} = \{\{(\boldsymbol{x}^u, \boldsymbol{x}^v) \mid (u, v) \in E\}\}$. Together with the order statistics, we now define the attribute hypergraph as follows:

Definition 3.1 (Attribute hypergraph $\mathcal{F}_\mathcal{E}$)**.** Given a multiset of attributes of the endpoint nodes $\mathcal{E} = \{\{(\boldsymbol{x}^u, \boldsymbol{x}^v) \mid (u, v) \in E\}\}$ of m entities with totally ordered attribute spaces, the feature hypergraph $\mathcal{F}_\mathcal{E}$ is defined as follows. First, we label the graph with m. Then,

- For each order statistic $\boldsymbol{x}_{i(k)}$ of attribute i and order k $(1 \leq k \leq m)$, there are 2 nodes, namely $(i, k, 1)$ and $(i, k, 2)$. In total, there are exactly $2md$ nodes in $\mathcal{F}_\mathcal{E}$ (attribute values need not be unique). Nodes $(i, k, 1)$ and $(i, k, 2)$ store a single attribute to mark their order: k.
- Let $o_i(u)$ be the order of the attribute value $\boldsymbol{x}_i^u$, i.e., $\boldsymbol{x}_{i(o_i(u))} = \boldsymbol{x}_i^u$. For each pair of attributes of endpoint nodes $(\boldsymbol{x}^u, \boldsymbol{x}^v) \in \mathcal{E}$, there is a hyperedge H_{uv} in $\mathcal{F}_\mathcal{E}$ defined as

$$H_{uv} \coloneqq \{(1, o_1(u), 1), (1, o_1(v), 2), (2, o_2(u), 1), (2, o_2(v), 2), \ldots, (d, o_d(u), 1), (d, o_d(v), 2)\}. \quad (4)$$

Our first observation is that the feature hypergraph in Definition 3.1 perfectly captures the order statistics of the set $\mathcal{E}$ but discards the actual values of the attributes.

We now consider statistical tests that measure dependencies of the attributes of endpoint nodes. As an example, consider that if $(\boldsymbol{x}^u, \boldsymbol{x}^v) \in \mathcal{E}$ are samples (not necessarily independently sampled) from a bivariate distribution $(\mathbf{x}, \mathbf{x}') \sim F$, one may be interested in testing the hypothesis

$$H_0 : F(\mathbf{x}, \mathbf{x}') = F_1(\mathbf{x}) F_2(\mathbf{x}'),$$

i.e., that $\mathbf{x}$ and $\mathbf{x}'$ are independent. Bell (1964); Berk & Bickel (1968) showed that over totally ordered sets, measures (e.g., p-values) of such hypothesis tests for pairwise independence (H_0 above) and higher-order conditional independence between multiple variables, have invariances that simplify the data representation to such a degree that the original values are discarded, retaining only the order relationships between the variable values. Any such test is therefore a rank test, i.e., it relies only on indices of the order statistic, not on the numerical values of the attributes.

Our first theoretical contribution is the observation that any statistical test that focuses on measuring the (conditional) dependencies of attributes of endpoint nodes in $\mathcal{E}$ can be defined as a graph regression task over the feature hypergraph $\mathcal{F}_{\mathcal{E}}$ of Definition 3.1.

Theorem 3.2. *Given a multiset of attributes of the endpoint nodes $\mathcal{E}$, the corresponding feature hypergraph $\mathcal{F}_{\mathcal{E}}$ (Definition 3.1) and a most-expressive hypergraph GNN encoder $M_{\theta^*}(\mathcal{F}_{\mathcal{E}})$, then any test $T(\mathcal{E})$ that focuses on measuring the dependence of the attributes of the endpoint nodes of $\mathcal{E}$ has an equivalent function h within the space of Multilayer Perceptrons (MLPs) that depends solely on the graph representation $M_{\theta^*}(\mathcal{F}_{\mathcal{E}})$, i.e., $\exists h \in$ MLPs s.t. $T(\mathcal{E}) = h(M_{\theta^*}(\mathcal{F}_{\mathcal{E}}))$.*

Next we show that the hypergraph $\mathcal{F}_{\mathcal{E}}$ can be simplified with STAGE-edge-graph and that the ability to compute dependency measures can be made invariant to certain domain shifts between train and test.

3.2. Transferability of STAGE

The feature hypergraph $\mathcal{F}_{\mathcal{E}}$ in Definition 3.1 is used to obtain a maximal invariant graph representation via hypergraph GNN. This solution has a high computational cost from the use of hypergraph GNNs. Fortunately, we show that by assigning unique attribute identifiers to label the nodes of our STAGE-edge-graphs $\boldsymbol{G}(\boldsymbol{S}^{uv})$ (Definition 2.1), STAGE-edge-graphs are as informative as the corresponding feature hypergraphs, preserve the same invariances, while allowing the usage of (non-hypergraph) GNN encoders.

Theorem 3.3. *Given the attributes of the endpoint nodes $\mathcal{E}$ (Definition 3.1) of a graph $G = (V, E, \boldsymbol{X})$, there exists an optimal parameterization θ_g^*, θ_s^* for a most expressive GNN encoder M^g and a most-expressive multiset encoder M^s, respectively, such that $M_{\theta_s^*,\theta_g^*}(G) := M^s_{\theta_s^*}\left(\left\{\left\{M^g_{\theta_g^*}(\boldsymbol{G}(\boldsymbol{S}^{uv})) : (u,v) \in E\right\}\right\}\right)$ such that any test $T(\mathcal{E})$ that measures the dependence of $\mathcal{E}$'s attributes of the endpoint nodes has an equivalent function h within the space of Multilayer Perceptrons (MLPs) that depends solely on the graph representation $M_{\theta_s^*,\theta_g^*}(G)$, i.e., $\exists h \in$ MLPs s.t. $T(\mathcal{E}) = h(M_{\theta_s^*,\theta_g^*}(G))$.*

Theorem 3.3 motivates the design of STAGE, which leverages a GNN on STAGE-edge-graphs to obtain edge-level embeddings. However, the use of unique attribute identifiers in the STAGE-edge-graphs disrupts invariance to permutations in attribute order (e.g., U.S. shoe size appearing as the first attribute in one dataset and U.K. shoe size as the last attribute in another), thereby limiting its domain transferability. More broadly, we now describe all the invariances we want for STAGE to have in order to be robust to a class of attribute domain shifts.

COGG invariances. STAGE-edge-graphs facilitate domain transfer to distinct attribute domains. Intuitively, the full set of invariances required for domain transferability over $G = (V, E, \boldsymbol{X})$ consists of: (1) invariance or equivariance to transformations of attribute values that preserve the order statistic, (2) invariance or equivariance to permutations of attribute orders (columns of $\boldsymbol{X}$), and (3) invariance or equivariance to permutations of nodes in the graph, affecting V (and consequently E) and the rows of $\boldsymbol{X}$. These invariances are formalized in Definition B.5 in Appendix B.4 through the actions of *component-wise order-preserving groupoid for graphs* (COGG). Importantly, groups are insufficient to capture these invariances because they assume transformations act within a single attribute domain. However, we are interested in transformations across distinct attribute spaces. Groupoids generalize groups by allowing these transformations between different domains, making them the natural choice for modeling the required invariances.

We now introduce our final theoretical contribution which establishes that STAGE achieves invariance to COGGs by design. This result shows that STAGE can provably achieve the zero-shot transferability to the class of attribute domain shifts defined by COGGs-type transformations.

Theorem 3.4. *STAGE is invariant to COGGs (Definition B.5).*

The proof sketch is as follows. From Theorem 3.3, STAGE achieves invariance to changes in attribute values, including their types, names, and semantics. Then, by dropping the attribute identifiers in STAGE-edge-graphs, we sacrifice maximal expressivity but ensure that STAGE is invariant to permutations of the attribute order. Finally, since STAGE employs a second GNN on the original input graph, using the embeddings of the STAGE-edge-graphs, while omitting the original node attributes, STAGE achieves invariance to node permutations. Thus, the method is invariant to COGGs.

4. Experiments

We demonstrate the effectiveness of STAGE across multiple experimental settings, focusing on small to medium-sized datasets. While the computational complexity scales linearly with the graph size and quadratically with the number of attributes, training on these datasets introduces only mod-

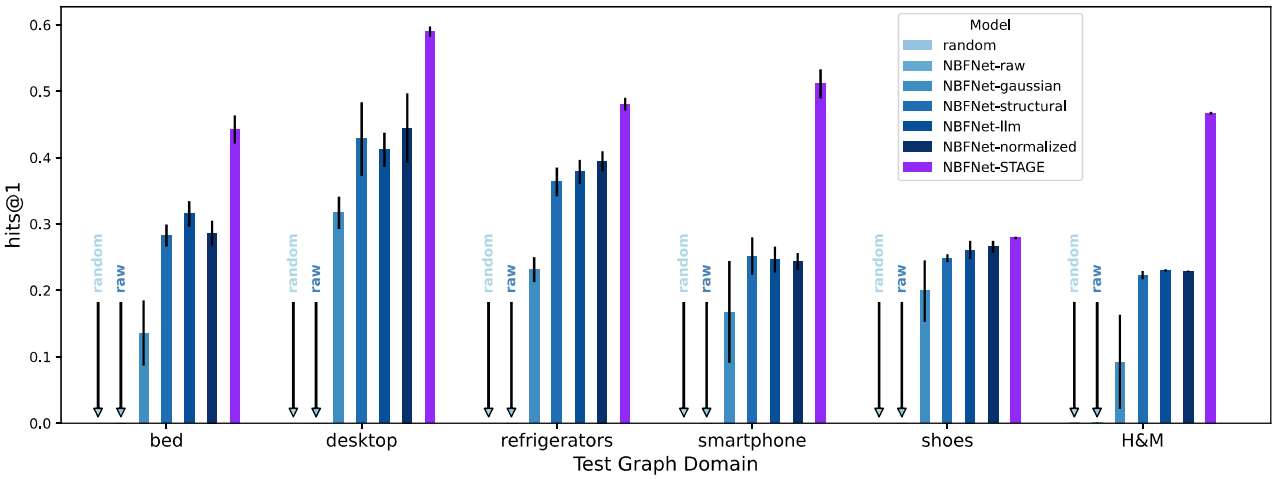


Figure 3: Zero-shot Hits@1 performance (higher is better) of STAGE and baselines, trained on four (or five) distinct E-Commerce Store domains and evaluated on the held-out domain (or H&M dataset). **NBFNet-STAGE consistently achieves the highest zero-shot accuracy across all test domains, with up to 103% improvement**.

erate computational overhead (e.g., 7.83% slower than the fastest baseline in link prediction; see Appendix H). Therefore, STAGE is highly effective in these settings, achieving strong generalization performance. In the following, we present our main results and refer to Appendix D for details. Our code is available at `https://github.com/snap-stanford/stage-gnn/`.

Datasets. To evaluate zero-shot generalization to graphs with new attributes, we consider datasets with distinct domain-specific attributes but a shared task. Our datasets contain graphs with up to 4k nodes, 50k edges, 16 attributes, representing small to medium-size real-world scenarios where STAGE is particularly effective. Due to space constraints, we introduce them below and refer to Appendix C.

E-Commerce Stores dataset (link prediction). We use data from a multi-category store (Kechinov, 2020) containing customer-product interactions (purchases, cart additions, views) over time. To simulate distinct single-category retailers, we partition the dataset into five domains, each representing a specialized store: *shoes*, *refrigerators*, *desktops*, *smartphones*, and *beds*. Each domain has its own customer base and product-specific attributes (e.g., *smartphones* have *display type*; *shoes* have *ankle height*). The task is to predict future customer-product interactions from past actions.

H&M dataset (link prediction). We use the H&M Personalized Fashion Recommendations dataset (Kaggle, 2021), which contains transactions from a large fashion retailer, to evaluate the zero-shot performance of models trained on E-Commerce Stores. All attributes, except for "price", differ from those in E-Commerce Stores. The task remains to predict customer-product interactions from past actions.

Social network datasets (node classification): Friendster and Pokec. We evaluate STAGE on two online social networks from different regions and user bases: Friendster (Teixeira et al., 2019) and Pokec (SNAP, 2012). Friendster nodes have attributes such as *age*, *gender*, *interests*, while Pokec nodes have *public profile status*, *completion percentage*, *region*, *age*, and *gender*. The task is to predict a node attribute common to both social networks using network structure and remaining node attributes. Since only *age* and *gender* are shared, we create two tasks: mask and predict *gender* (presented in this section), and mask and regress on *age* (discussed in Appendix E).

Baselines. We compare STAGE to several baselines designed to handle new node attributes. (1) *raw*: Projects each raw node attribute into a fixed-dimensional space with a linear transformation, before summing across the projected dimensions. (2) *gaussian*: Use Gaussian noise as node attributes (Sato et al., 2021; Abboud et al., 2021). (3) *structural*: Ignores node attributes entirely, using only the graph structure. (4) *llm*: Converts node attributes into textual descriptions and obtains embeddings using a pretrained encoder-only language model, taking only the node attributes as input (without graph structure) due to prompt length limitations, similar to PRODIGY (Huang et al., 2023). (5) *normalized*: Retains only continuous attributes and standardize them. For a fair comparison, all methods utilize the same underlying GNN architecture, NBFNet (Zhu et al., 2021c) for link prediction and GINE (Hu et al., 2020) for node classification. In Appendix F, we report additional experiments with other architectures. In addition to these baselines, we evaluate our approach against *GraphAny* (Zhao et al., 2024b), a recent method for domain transferability in node

classification tasks, but not applicable to link prediction.

4.1. Zero-Shot Link Prediction on Unseen Domains

We evaluate the performance of all methods on zero-shot generalization on the E-Commerce Stores dataset, training on four categories, and testing on the held-out fifth category.

Results. As shown in Figure 3, STAGE consistently outperforms all baselines in zero-shot Hits@1 across all test domains. Notable improvements include: 103% gain when testing on the *smartphone* category (0.51 vs 0.25 Hits@1), 40% on *bed* (0.44 vs 0.31), and 33% on *desktop* (0.59 vs 0.44) compared to the strongest baselines.

In Table 1, we report the average performance of each model, calculated by taking the results in which each domain is held out once and averaging the scores. Our evaluation also includes popular non-parametric link prediction approaches such as Common Neighbors, Adamic Adar, and Personalized PageRank, with results showing that STAGE substantially outperforms classical heuristic methods by 54%, 51%, and 3837% respectively on Hits@1, while maintaining similar performance advantages on MRR. Overall, STAGE achieves 41% higher average Hits@1 (0.46 vs 0.33) and 29% higher MRR (0.50 vs 0.38) against the strongest baseline (normalized), with lower variance across seeds. This emphasizes the benefit of STAGE in transforming node attributes into a *unified input space* using learned edge embedding via *STAGE-edge-graph*, including its stronger attribute representation capabilities than LLM-based encoding approaches in the medium-sized graphs considered in this work.

4.2. Cross-Dataset Zero-Shot Link Prediction

We evaluate models trained on E-Commerce Stores for zero-shot prediction on the H&M dataset, which has distinct customers, products, activity patterns and attributes.

Table 1 shows that the performance on H&M of STAGE when trained on E-Commerce Stores is virtually identical to its performance on the held-out category in E-Commerce Stores (0.46 vs. 0.46 Hits@1). This highlights the robustness of STAGE to domain shifts, as it maintains similar performance when transitioning from E-Commerce Stores, which primarily feature household items, electronics, and shoes, to H&M, which focuses on clothing with minimal overlap in product types.

In Hits@1, STAGE achieves a relative improvement of 103% over the best parametric baseline (*llm*) (0.46 vs. 0.23). Moreover, STAGE obtains a relative improvement of 202% against a supervised structural method trained and tested on H&M (*structural-supervised*). In MRR, STAGE achieves the highest score, outperforming the best baseline by 99%.

Moreover, STAGE demonstrates a substantial improvement of 99% in Hits@1 over Adamic Adar (0.466 vs 0.2349), which performs the best among traditional heuristic methods on the H&M dataset. Similarly, STAGE outperforms Adamic Adar by 48% in MRR (0.4703 vs. 0.3184), further confirming the its superiority over classical link prediction heuristics in zero-shot scenarios.

4.3. Zero-Shot Node Classification on Unseen Domains

To validate our approach beyond link prediction and E-Commerce scenarios, we benchmark on a node classification task using two social network datasets, where the goal is to predict user *gender*. We train models on Friendster and evaluate zero-shot on Pokec.

Table 2 shows that STAGE achieves a 10.3% improvement over the best baseline (and lower variance), also surpassing the task-specific model GraphAny (Zhao et al., 2024b) and the cross-domain pretraining method GCOPE (Zhao et al., 2024a). This indicates that STAGE effectively captures attribute dependencies also in node classification tasks and outperforms all approaches by leveraging its unified input space obtained by the usage of the STAGE-edge-graphs.

4.4. Generalization When Training on Multiple Domains

We examine how the model performance varies with the number of training domains in E-Commerce Stores.

As shown in Figure 4, STAGE obtains improving zero-shot performance (both Hits@1 and MRR) with more training domains. While not the only method showing improvement, STAGE exhibits notably tighter interquartile ranges compared to the only other method exhibiting better performance with increasing domain (*gaussian*) at higher domain counts. Additionally, STAGE's lower whiskers consistently rise with more domains, showing also that its worst-case scenarios improve with more training data.

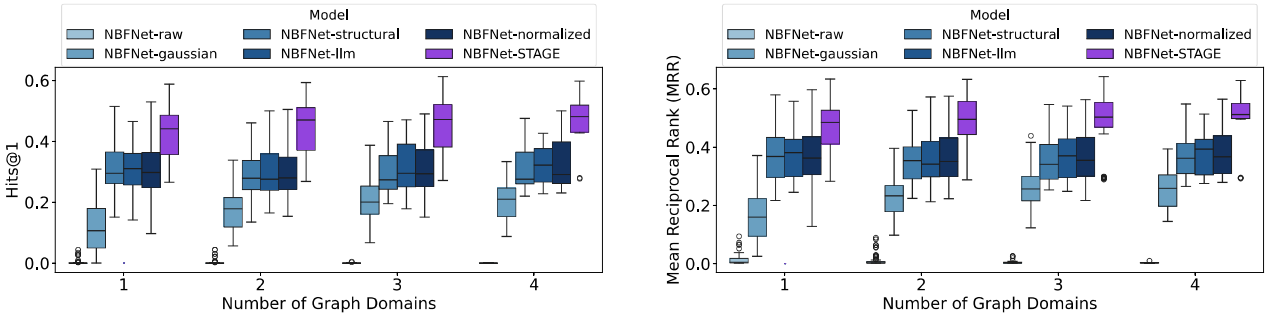
These results further validate that STAGE is capable of learning transferable patterns across domains through its defined *unified input space*. The consistent performance gains with additional training domains suggest that *STAGE-edge-graph* effectively captures generalizable dependencies between attributes, with more training domains enabling the learning of a broader range of dependencies. In contrast, baseline approaches that ignore attributes or use generic embeddings fail to leverage the additional training domains for improved cross-domain generalization.

5. Related Work

In this section, we present the most closely related works to our STAGE. A more in-depth comparison, along with additional related work, can be found in Appendix I.

Table 1: **NBFNet-STAGE outperforms all baselines in zero-shot Hits@1 and MRR (including supervised approaches) across the E-Commerce Stores and H&M datasets.** For the E-Commerce Stores, results are averaged across models trained on all combinations of four graph domains and tested on the remaining domain. For zero-shot test on H&M, models are trained on the five E-Commerce Stores domains. % **gain** shows relative improvement of STAGE over each baseline.

Training: E-Commerce Stores	**Test: Held-out E-Comm. Store**				**Test: H&M Dataset**			
Model	**Hits@1** (↑)	% **gain**	**MRR**	% **gain**	**Hits@1** (↑)	% **gain**	**MRR** (↑)	% **gain**
random	0.0026 ± 0.0000	17615%	-	-	0.0006 ± 0.0000	77667%	-	-
Common Neighbors	0.2991 ± 0.0006	54%	0.3942 ± 0.0014	26%	0.2354 ± 0.0000	98%	0.3179 ± 0.0000	48%
Adamic Adar	0.3052 ± 0.0007	51%	0.4001 ± 0.0015	24%	0.2349 ± 0.0000	99%	0.3184 ± 0.0000	48%
Personalized PageRank	0.0117 ± 0.0000	3837%	0.0714 ± 0.0001	596%	0.0105 ± 0.0000	4344%	0.0717 ± 0.0000	556%
NBFNet-raw	0.0000 ± 0.0000	∞	0.0032 ± 0.0009	15434%	0.0005 ± 0.0004	93220%	0.0059 ± 0.0011	7871%
NBFNet-gaussian	0.2101 ± 0.0428	119%	0.2617 ± 0.0459	90%	0.0925 ± 0.0708	404%	0.1176 ± 0.0756	300%
NBFNet-structural	0.3149 ± 0.0253	46%	0.3721 ± 0.0219	34%	0.2231 ± 0.0060	109%	0.2302 ± 0.0080	104%
NBFNet-llm	0.3226 ± 0.0190	43%	0.3830 ± 0.0145	30%	0.2302 ± 0.0015	103%	0.2365 ± 0.0021	99%
NBFNet-normalized	0.3269 ± 0.0213	41%	0.3844 ± 0.0159	29%	0.2286 ± 0.0010	104%	0.2341 ± 0.0018	101%
NBFNet-structural-supervised	N/A	N/A	N/A	N/A	0.1546 ± 0.0084	202%	0.2103 ± 0.0164	124%
NBFNet-STAGE (Ours)	**0.4606** ± 0.0123	0%	**0.4971** ± 0.0073	0%	**0.4666** ± 0.0020	0%	**0.4703** ± 0.0029	0%



(a) Zero-shot Hits@1

(b) Zero-shot MRR

Figure 4: **The performance (both Hits@1 and MRR) of STAGE improves with more train domains, while this is not the case for other methods.** Box-plot distribution over all combinations of a fixed number of graph domains in the E-Commerce Stores dataset and testing on the held-out domain(s), averaged over random seeds.

Graphs Generalization under Distribution Shifts. Several works address distribution shifts between train and test graphs over the same attribute domain, such as Shen et al. (2023); Zhu et al. (2021b), which employ learned augmentations to mitigate the change in distribution in test. Meanwhile, extensive research has focused on domain adaptation for GNNs (Dai et al., 2022; Li et al., 2020; Kong et al., 2022; Pei et al., 2020; Veličković et al., 2019; Wiles et al., 2022; Zhang et al., 2019; Zhu et al., 2021a), which typically assume access to data in both source and target domains. In contrast, our work tackles the more challenging scenario of zero-shot generalization to unseen attribute domains. To the best of our knowledge, all out-of-distribution graph methods (Zhang et al., 2024a) do not address the attribute domain shifts we consider, which include changes in the number of attributes between train and test.

Foundation Models for Graphs. Developing foundation models for graph data is a growing research interest, aiming to create versatile graph models capable of generalizing across different graphs and tasks (Mao et al., 2024). Initial efforts in this direction convert attributed graphs into texts and apply an LLM (Liu et al., 2024; Chen et al., 2024b;a; Tang et al., 2024; Zhao et al., 2023; He & Hooi, 2024; Huang et al., 2023). However, while promising, this methodology risks information loss and may limit transferability (Collins et al., 2024; Gruver et al., 2024; Schwartz et al., 2024). In contrast, non-LLM approaches attempt to directly address domain transferability in the attribute space (Xia & Huang, 2024; Lachi et al., 2024; Zhao et al., 2024b; Frasca et al., 2024; Yu et al., 2024; Zhao et al., 2024a), or by avoiding the use of node attributes entirely (Gao et al., 2023; Lee et al., 2023; Galkin et al., 2024; Zhang et al., 2024b). We provide details in Appendix I about why some of these approaches are not applicable as our baselines.

Table 2: Zero-shot test accuracy of STAGE and baselines on the Pokec dataset, trained on Friendster. % **gain** shows relative improvement of STAGE over each baseline.

Model	**Accuracy** (↑)	**% gain**
random	0.500 ± 0.0000	30.4%
GINE-raw	0.558 ± 0.0829	16.8%
GINE-gaussian	0.588 ± 0.0250	10.9%
GINE-structural	0.564 ± 0.0466	15.6%
GINE-llm	0.550 ± 0.0368	18.5%
GINE-normalized	0.541 ± 0.0148	20.5%
GraphAny	0.591 ± 0.0083	10.3%
GCOPE	0.535 ± 0.0153	21.9%
GINE-STAGE (Ours)	**0.652 ± 0.0042**	0%

6. Conclusion and Future Work

The challenge of learning universal graph representations that generalize across diverse attribute domains has limited progress in graph foundation models, mainly due to the lack of a unified input space to represent node attributes, which may vary in test graphs. In this paper, we proposed STAGE, which addresses this limitation by transforming diverse attribute spaces into a unified representation, learning statistical dependencies between attributes instead of relying on their absolute values. By demonstrating that these dependencies remain invariant under certain domain shifts, STAGE provides theoretical foundations for zero-shot generalization across graphs with differing attribute spaces. Our strong empirical results on medium-sized datasets demonstrate the practical effectiveness of this approach.

While STAGE represents a meaningful step forward, it also highlights opportunities for future research. The unified input space we introduce could serve as a basis for developing graph foundation models that can learn from diverse graph datasets at scale, reducing the quadratic complexity of STAGE. However, realizing this potential will require addressing additional challenges, such as developing architectures to capture complex high-order attribute dependencies and scaling to large graph collections.

Acknowledgments

BR and BB acknowledge support from the National Science Foundation (NSF) awards CCF-1918483, CAREER IIS-1943364 and CNS-2212160, an Amazon Research Award, and AnalytiXIN, Wabash Heartland Innovation Network (WHIN), Ford, NVidia, CISCO, and Amazon. Computing infrastructure was supported in part by CNS-1925001 (CloudBank). This work was supported in part by AMD under the AMD HPC Fund program.

JL gratefully acknowledges the support of NSF under Nos. OAC-1835598 (CINES), CCF-1918940 (Expeditions), DMS-2327709 (IHBEM), IIS-2403318 (III); Stanford Data Applications Initiative, Wu Tsai Neurosciences Institute, Stanford Institute for Human-Centered AI, Chan Zuckerberg Initiative, Amazon, Genentech, GSK, Hitachi, SAP, and UCB. The content is solely the responsibility of the authors and does not necessarily represent the official views of the funding entities.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Abboud, R., Ceylan, İ. İ., Grohe, M., and Lukasiewicz, T. The surprising power of graph neural networks with random node initialization. In *Proceedings of the Thirtieth International Joint Conference on Artifical Intelligence (IJCAI)*, 2021.

Bell, C. B. A characterization of multisample distribution-free statistics. *Annals of Mathematical Statistics*, 35(2): 735–738, 1964. doi: 10.1214/aoms/1177703564.

Berk, R. and Bickel, P. On invariance and almost invariance. *Annals of Mathematical Statistics*, 39(5):1573–1576, 1968. doi: 10.1214/aoms/1177698328.

Berk, R., Nogales, A., and Oyola, J. Some counterexamples concerning sufficiency and invariance. *The Annals of Statistics*, pp. 902–905, 1996.

Berk, R. H. A remark on almost invariance. *The Annals of Mathematical Statistics*, pp. 733–735, 1970.

Bevilacqua, B., Robinson, J., Leskovec, J., and Ribeiro, B. Holographic node representations: Pre-training task-agnostic node embeddings. In *The Thirteenth International Conference on Learning Representations*, 2025.

Chen, R., Zhao, T., Jaiswal, A. K., Shah, N., and Wang, Z. LLaGA: Large language and graph assistant. In *Forty-first International Conference on Machine Learning*, 2024a.

Chen, Z., Mao, H., Li, H., Jin, W., Wen, H., Wei, X., Wang, S., Yin, D., Fan, W., Liu, H., and Tang, J. Exploring the potential of large language models (llms) in learning on graphs. *ACM SIGKDD Explorations Newsletter*, 25(2): 42–61, 2024b.

Collins, K. M., Jiang, A. Q., Frieder, S., Wong, L., Zilka, M., Bhatt, U., Lukasiewicz, T., Wu, Y., Tenenbaum, J. B., Hart, W., et al. Evaluating language models for mathematics through interactions. *Proceedings of the National Academy of Sciences*, 121(24):e2318124121, 2024.