Towards Locality-Aware Meta-Learning of Tail Node Embeddings on Networks

Zemin Liu1∗, Wentao Zhang2∗†, Yuan Fang1, Xinning Zhang2, Steven C.H. Hoi1,3‡
1Singapore Management University, 2University of Science and Technology of China, 3Salesforce Research Asia
{zmliu, yfang}@smu.edu.sg, zwt@mail.ustc.edu.cn, xinning@ustc.edu.cn, shoi@salesforce.com

ABSTRACT

Network embedding is an active research area due to the prevalence of network-structured data. While the state-of-the-art often learns high-quality embedding vectors for structure-rich nodes (i.e., nodes with a sufficient number of links), the quality of the embedding vectors for low-degree or tail nodes is often suboptimal due to the lack of structural information. To date, little efforts have been devoted to learning tail node embeddings, which is especially important as many real-world networks are long-tailed. In this paper, we formulate the goal of learning tail node embeddings into a few-shot regression problem, given the very small number of links on the tail nodes. Furthermore, since each node resides in its own local context, we personalize the way of learning the embedding vector for each tail node. Due to the extreme sparse structural information carried by the tail nodes, node personalization would present severe overfitting. To overcome the challenge, we propose a locality-aware meta-learning framework, called meta-tail2vec, which learns to embed the embedding model for the tail nodes at different localities. Finally, we conduct extensive experiments and demonstrate the promising results of meta-tail2vec. (Supplemental materials including code and data are available at https://github.com/smufang/meta-tail2vec.)

CCS CONCEPTS
• Information systems → Data mining; • Computing methodologies → Machine learning: Learning latent representations.

KEYWORDS
meta-learning, network embedding, tail nodes

ACM Reference Format:

∗Co-first authors with equal contribution.
†Work done as a visiting research student at Singapore Management University.
‡Currently on leave from Singapore Management University.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

CIKM ’20, October 19–23, 2020, Virtual Event, Ireland
© 2020 Association for Computing Machinery.
ACM ISBN 978-1-4503-6859-9/20/10. . $15.00
https://doi.org/10.1145/3340531.3411910

Figure 1: Distribution of node degree and its relationship to the quality of embedding vector on the Wiki network.

1 INTRODUCTION

Network structures are prevalent in various real-world scenarios, such as social networks, citation networks and biological networks. On these networks, many problems can be formulated as node classification and link prediction, which rely heavily on effective network representations. While traditional approaches mainly focus on manual feature engineering, recent network embedding [26] and graph neural networks [17] have become the de facto state-of-the-art approaches for learning representations of network data. Specifically, these methods aim to encode network structures by mapping nodes into a low-dimensional vector space, in which the structural information is preserved.

Despite their success, a critical question remains open: the performance of most existing embedding methods depend on the availability of abundant structural information. Although this might not be a concern for medium-to-high degree nodes with many links to other nodes, low-degree nodes with very few links tend to suffer from the problem of limited structural information. That is, the learned embedding vector for a low-degree node cannot accurately model its structural information, resulting in suboptimal performance. Generally, the node degrees vary considerably across the network and are not uniformly distributed. In many networks, the degrees approximately follow the power-law distribution [26]. For instance, Fig. 1(a) illustrates the degree distribution of the Wiki network [39], a network of interlinked Wikipedia pages belonging to 19 categories. The node degrees are characterized by a long-tailed distribution, where a significant fraction of nodes belong to the tail with very low degrees. The embeddings of these tail nodes are unsatisfactory, as demonstrated by the performance of node classification in Fig. 1(b). Particularly, when the node degree increases, the performance also increases, due to the observation of more structural information.

Unfortunately, most current network embedding or representation learning approaches overlook the tail nodes, by regarding all...
nodes as the same and adopting the same learning approach despite their diverse degrees. For example, DeepWalk [26] samples node sequences from the network and feeds them to the same skip-gram model without paying special attention to low-degree nodes, while graph neural networks [12, 17] derive the embedding of a target node by aggregating its neighboring nodes in the same manner without adapting to the degree of the target node. Not surprisingly, the tail nodes with very few links are usually under-modeled than those with many links. This inspires us to investigate the following research problem: How do we learn effective embedding vectors for tail nodes from limited structural information?

The problem is challenging for two reasons. (1) The tail nodes have very few links, which provide scarce structural information. The issue is especially severe when only network structures are available, without access to additional side information (e.g., content for items in recommendation [36], or morphological information in word embedding [1]). (2) Each node is associated with a unique locality, and thus the embedding model should ideally become specialized from node to node. However, in practice this presents a dilemma: adapting to each node often causes overfitting, which is particularly serious for tail nodes with very few observed links.

To address the first challenge of scarce structural information, we exploit the embeddings of nodes with a sufficient number of links (i.e., non-tail nodes and we call them head nodes hereafter). Considering any network embedding model, the learned embedding vectors of head nodes tend to be more accurate than those of tail nodes, due to more abundant structural information associated with head nodes. Thus, we treat the embeddings of head nodes as our oracle embeddings, which can be leveraged to train a regression model capable of reconstructing the oracle embeddings in a self-supervised manner. Using the regression model, our goal is to predict new embeddings for tail nodes such that their quality can approach that of the oracle embeddings. To further ensure that the regression model fits both head and tail nodes, we propose to perform link dropouts on head nodes, to simulate the limited structural information of tail nodes.

To address the second challenge of adapting to the locality of each node, we resort to the meta-learning paradigm [9, 28], for its ability of adapting to new learning tasks (also called episodes) by learning a knowledge prior common to different tasks. Known as meta-learning, it learns how to learn a model in the form of a prior, contrary to learning one model for all tasks (which cannot differentiate tasks) or individual models for each task (which may overfit to each task). In particular, we adopt the meta-learning framework MAML [9], which has the advantage of quick adaptation in a model-agnostic manner with considerable success in various problem domains [15, 19].

To realize the above insights in our scenario, we construct locality-aware tasks: one task for each node to represent its locality, and learn a regression prior that can be easily adapted to each task or locality. In each task, given a query node, we aim to predict its embedding vector after training on a set of support nodes. While the support set is often randomly sampled, for locality-aware learning, we propose to use the neighbors of the query nodes as the support nodes, assuming that the localities of neighboring nodes are similar. More specifically, during meta-training, each task involves a head node as the query, with the objective of learning a regression prior such that the prior can be adapted by the support nodes to accurately reconstruct the oracle embeddings of the query. During meta-testing, each task involves a tail node as the query, and the learned regression prior will be adapted by the support nodes to predict a better embedding vector for the query. To make the meta-training and meta-testing tasks more similar, we again adopt link dropout to sample only a few neighbors as the support set during meta-training, to simulate meta-testing tasks on tail nodes. Essentially, each task is a few-shot regression problem.

In summary, we propose a novel approach meta-tail2vec, which learns to learn tail node embeddings in a few-shot regression setting. The model operates in an embedding-agnostic manner, which is flexible to work with any network embedding model. Specifically, our contribution is threefold. (1) We formulate the under-explored problem of learning tail node embedding on networks, and cast it into an instance of regression via oracle reconstruction. (2) We propose a base regression model hinged on the concept of link dropouts, and formulate the locality-aware tasks on networks in a meta-learning framework, which allows for easy local adaption of the base model. (3) We conduct extensive experiments on three public datasets, in which meta-tail2vec attains significant performance gains.

2 RELATED WORK

Network embedding [2] has been popular for graph representation learning, requiring only network structures as input. Various approaches have been proposed to generate structure-preserving node embeddings, such as random walks [11, 26], matrix factorization [4, 23], node proximity [31], and motifs or subgraphs [7, 38]. Graph neural networks (GNN) [12, 17] emerge as another powerful tool for representation learning, which are designed to integrate both node features and network structures.

However, little attention has been paid to the learning of tail node embeddings. While several recent studies explore various techniques to deal with sparse networks, such as using dual dropouts to avoid overfitting to sparse links [3], adopting the adversarial principle to learn the underlying distribution [25, 35] and leveraging the global network structure [34], they aim to increase the overall robustness of the learning process, and do not specifically target the most vulnerable tail nodes. Nevertheless, some studies have been devoted to related problems with sparse observations in other domains, such as cold-start recommendation [20, 33, 36] and rare word embedding [1, 15, 27]. However, these methods rely on additional side information, such as item contents or word morphology.

To address the general problem of learning from less data, particularly in few-shot scenarios, meta-learning [9, 29, 32] has demonstrated considerable success in several domains including vision [32], language [13, 15] and data mining [19, 24]. These methods typically learn some prior knowledge from an abundant number of related tasks, and adapt the prior to new tasks with limited data. As many forms of data often entail inherent graph structures, recent studies often exploit some auxiliary graphs, such as affinity graphs [10, 22], class relational graphs [21] and one-hop ego-networks [37, 42]. Few-shot learning on graphs proper have also been explored, including node classification on a single large graph [43], and node or graph-level classification on multiple graphs [5, 40]. To the best of our knowledge, our approach is the first use of meta-learning for tail node embedding.
3 PROBLEM CASTING AS REGRESSION

We begin with the problem statement, and cast it as an instance of regression based on the idea of oracle reconstruction.

3.1 Problem statement

Consider a graph \( G = (\mathcal{V}, E) \) with the set of nodes \( \mathcal{V} \) and the set of edges \( E \). Let \( \mathcal{N}_v \) denote the set of neighbors of node \( v \in \mathcal{V} \), and the size of the neighbor set \( |\mathcal{N}_v| \) is known as the degree of \( v \). In particular, \( v \) is a low-degree or tail node if its degree is no larger than some constant \( k \). That is, the set of tail nodes is \( \mathcal{V}_{\text{tail}} = \{ v \in \mathcal{V} : |\mathcal{N}_v| \leq k \} \). We call the remaining nodes head nodes: \( \mathcal{V}_{\text{head}} = \{ v \in \mathcal{V} : |\mathcal{N}_v| > k \} \).

Given any base network embedding model \( \phi \), we denote \( h_v = \phi(G, v) \in \mathbb{R}^d \) as the \( d \)-dimensional embedding vector of node \( v \), learned by \( \phi \) on graph \( G \), \( \forall v \in \mathcal{V} \). Since the tail nodes have scarce structural features due to their small degrees, presumably their embeddings \( \{ h_v : v \in \mathcal{V}_{\text{tail}} \} \) are inferior in quality to the embeddings of the head nodes \( \bar{\mathcal{O}} = \{ h_v : v \in \mathcal{V}_{\text{head}} \} \).

Our goal is to learn new embedding vectors for the tail nodes \( \{ h_v : v \in \mathcal{V}_{\text{tail}} \} \), such that their quality is improved to eventually approach the quality of the head nodes’ embeddings \( \bar{\mathcal{O}} \). Note that our setup is embedding-agnostic, i.e., any embedding model \( \phi \) can be used as the base embedding model.

3.2 Casting as regression via oracle reconstruction

A major challenge of learning tail node embeddings lies in the limited availability of structural information. That is, each tail node only has a few observed neighbors. Without assuming additional side information (e.g., item contents in recommendation systems), we take advantage of the high-quality embeddings of head nodes, and investigate how high-quality embeddings can be similarly constructed for the tail nodes.

Specifically, we propose to cast the problem as an instance of regression. We train a regression model \( F \) on the head nodes, treating their embeddings \( \bar{\mathcal{O}} \), which are learned by the base embedding model \( \phi \), as the oracle embeddings w.r.t. \( \phi \). Given any head node \( v \), \( F(v; \Theta) \) outputs a new predicted embedding \( \hat{h}_v \) to approximate its oracle embedding \( h_v \in \mathcal{O} \). In other words, \( F \) is expected to reconstruct the oracle embeddings \( \mathcal{O} \) of the head nodes, regardless of how the base embedding model works. Formally, we optimize the parameters \( \Theta \) of the regression model \( F \) by

\[
\arg \min_{\Theta} \sum_{v \in \mathcal{V}_{\text{head}}} \| F(v; \Theta) - h_v \|^2. \tag{1}
\]

Note that here we use the Euclidean norm \( \| \cdot \| \), although other distance functions can be adopted too.

After training, \( F \) can be applied to the tail nodes as test instances, so as to predict their unknown oracle embeddings, supposedly attaining similar quality as \( \mathcal{O} \). The improved embedding vectors of tail nodes can be subsequently used as representations for downstream tasks such as node classification and link prediction, whose performance would benefit from the better embeddings.

4 META-LEARNED FEW-SHOT REGRESSION

In this section, we first introduce a regression model to predict new embedding vectors for tail nodes, then personalize the regression for each node by formulating node-wise locality-aware tasks, and finally present the overall meta-learning approach meta-tail2vec.

4.1 Embedding regression model

As illustrated in Fig. 2, we materialize the regression model \( F(v; \Theta) \) in order to reconstruct the oracle embedding \( h_v \) of a node \( v \). Specifically, we leverage the widely used Multi-Layer Perceptron (MLP):

\[
\hat{h}_v = F(v; \Theta) = W_2 \cdot \sigma(W_1 x_v + b_1) + b_2, \tag{2}
\]

where \( x_v \in \mathbb{R}^d \) is the input feature vector of node \( v \). The parameters of the MLP include \( W_1 \in \mathbb{R}^{d \times d_1}, W_2 \in \mathbb{R}^{d \times d_2}, b_1 \in \mathbb{R}^{d_1} \) and \( b_2 \in \mathbb{R}^d \), i.e., \( \Theta \equiv \{ W_1, W_2, b_1, b_2 \} \). \( \sigma(\cdot) \) is an activation function, and we adopt ReLU in this paper. Note that the size of the output layer is \( d \), the same as the embedding dimension of the nodes.

For a node \( v \), the input to the MLP is a feature vector \( x_v \). In the following, we discuss the formulation of the input feature vector.

4.1.1 Neighborhood aggregation. On a network, a node \( v \) is characterized by its structural features or its neighbor set, \( \mathcal{N}_v \). Naturally, we can aggregate the embedding vectors of the neighbors to construct its input feature \( x_v \), an idea similar and central to many graph neural networks [12, 17]. Specifically,

\[
x_v = \text{AGGR}(\{ h_i : i \in \mathcal{N}_v \}), \tag{3}
\]

where \( \text{AGGR}(\cdot) \) is an aggregator such as mean pooling. More generally, we can also aggregate nodes within an \( m \)-hop radius of \( v \), as shown in Fig. 2(a). Defining the \( m \)-hop neighbor set of \( v \) as

\[
\mathcal{N}_v^{(m)} = \bigcup_{i \in \mathcal{N}_v^{(m-1)}} \mathcal{N}_i, \tag{4}
\]

and \( \mathcal{N}_v^{(1)} \equiv \mathcal{N}_v \), the input feature \( x_v \) can be constructed as

\[
x_v = \text{AGGR}(\{ h_i : i \in \mathcal{N}_v^{(1)} \cup \mathcal{N}_v^{(2)} \cup \ldots \cup \mathcal{N}_v^{(m)} \}). \tag{5}
\]

Note that, although advanced strategies such as tree-LSTMs [30] and graph convolutions [17] may be employed to aggregate multiple hops, we adopt the simple mean pooling to demonstrate the effectiveness of our approach.
4.1.2 Link dropouts. One major flaw of such an input vector $x_v$ is that the head nodes in training and the tail nodes in testing possess very different neighbor sets in terms of their abundance, i.e., $|N_v| > |N_u|$ for some $v \in V_{\text{head}}$ and $u \in V_{\text{tail}}$. To make the training and testing nodes more similar, we perform link dropouts on the head nodes. Specifically, we randomly sample only $k$ neighbors of each head node for aggregation, in order to simulate the tail nodes. That is, given the sampled neighbors $\tilde{N}_v$ of a head node $v$, for one-hop aggregation we have

$$x_v = \text{AGGR}(\{h_i : i \in \tilde{N}_v\}), \quad \forall v \in V_{\text{head}}. \quad (6)$$

The idea of link dropouts draws an interesting parallel to Dropout-Net [33]. In their approach designed for cold-start recommendation, they eliminate the structural factor of users or items so as to simulate a cold-start scenario, forcing the model to fall back to user or item contents. However, we do not assume any sidewise information like user contents, which means our problem is more challenging and we can only drop the structural information partially.

4.2 Locality-aware few-shot regression tasks

When applying the regression model on all nodes, it ignores the unique locality of each node. As the nodes reside across different localities on the network, assuming one global model to fit all nodes is unrealistic. At the other extreme, learning an individual model for each node, including the popular pre-training and fine-tuning strategy [6], is likely to cause severe overfitting due to the limited structural information at the locality of each tail node.

4.2.1 Locality-aware support sets. To address the challenge of adapting to the locality of each node, we resort to the episodic meta-learning paradigm [9, 28]. The framework consists of many similar-natured learning tasks, divided into meta-training and meta-testing tasks. While each task is an instance in the meta-learning process, each task itself is a learning problem consisting of support and query sets (representing the usual sense of training and testing data, respectively). The goal of the meta-learning is to extract prior knowledge common to all tasks in meta-training, such that the knowledge can be quickly adapted to new tasks in meta-testing. In other words, it learns how to learn a task in the form of a common prior, instead of directly learning each task.

In our context, each task represents the unique locality of a node. At the task level, given a query node, we aim to predict its embedding vector after training on a set of support nodes. At the meta-learning level, we learn a prior regression model $F$ parameterized by $\Theta$ from the meta-training tasks where the query of each task is a head node, and further adapt the prior to the learning of new tasks in meta-testing where the query of each task is a tail node. However, in traditional meta-learning for classification tasks [9, 43], the support set for a query is randomly sampled. In such random sampling, the support set is not related to the query, and thus cannot reflect the unique locality of each query node. To generate locality-aware tasks, we propose to use the neighbors of the query node as the support nodes. The assumption is that the localities of neighboring nodes are similar, and thus training on these support nodes would be also applicable to the query node which lies in the vicinity of the support nodes. In other words, in each of our tasks, the support and query nodes are coupled based on their locality.

4.2.2 Formulation of few-shot tasks. As the neighbor sets of head and tail nodes differ vastly in size, a meta-training task with a head node as the query would have many support nodes, and a meta-testing task with a tail node as the query would have few support nodes. To make the tasks more similar, we again adopt link dropout by sampling only $k$ neighbors as the support set for meta-training tasks to simulate meta-testing tasks. Thus, each task becomes a few-shot (up to $k$ shots) regression problem, to predict the embedding vector of a query node from a few (up to $k$) support nodes.

We illustrate the task formulation with an example. As shown in Fig. 3(a), assuming $k = 3$, $v$ and $u$ is a head and tail node, respectively. On the one hand, the head node $v$ will be used to formulate a meta-training task: $v$ itself will be the query node, whereas we sample $k$ nodes from $v$’s neighbors $\tilde{N}_v = \{a, b, c, d, e\}$, to form the support set $\tilde{N}_v$, say, $|\tilde{N}_v| = \{a, b, c\}$. On the other hand, the tail node $u$ will be used to formulate a meta-testing task: $u$ itself will be the query node, whereas we simply take all of $u$’s neighbors $\tilde{N}_u = \{x, y, z\}$ as the support set. More example tasks are also illustrated in Fig. 3(b).
We employ MAML [9] for the meta-learning of tail node embeddings, which is capable of learning a prior $\Theta$ for any model using gradient-based optimization. In our case, the prior is the embedding regression model $F$, parameterized by $\Theta$. Different from the simple pre-training of a model, the prior $\Theta$ is learned in such a way that $\Theta$ can be quickly adapted to a new task by performing just one or a few gradient updates on the support set of the new task. The model $\Theta'$ adapted from the prior $\Theta$, is a local model for the query node in the same task.

More specifically, in our meta-training, consider a task $T_v = (S_v, q_v)$. As show in Fig. 4(b) and (c), the prior $\Theta$ itself is not directly updated or optimized by the support set $S_v$. Instead, it will be adapted by $S_v$ to produce a local model $\Theta'_v$ for the query $v$, through one or a few gradient updates w.r.t. $S_v$’s loss. The adapted local model $\Theta'_v$ will be applied to the query $v$ to predict an embedding vector $h_v$, so that the prior $\Theta$ can be updated by minimizing the task loss, i.e., the distance between the predicted embedding $h_v$ and the oracle embedding $h_0$ of the query node $v$.

Formally, let the loss of the prior on the support set $S_v$ be

$$L_{S_v}(\Theta) = \sum_{(i, h_i) \in S_v} ||F(i; \Theta) - h_i||^2.$$  \hspace{1cm} (7)

The prior $\Theta$ will be adapted by $S_v$ by one (or a few) gradient updates to generate a local model $\Theta'_v$ for the task $T_v$. That is,

$$\Theta'_v = \Theta - a \frac{\partial L_{S_v}(\Theta)}{\partial \Theta},$$  \hspace{1cm} (8)

where $a$ is the learning rate for the adaptation. Afterwards, the local model $\Theta'_v$ will be applied to the query $v$, to calculate the task loss using the query $q_v$:

$$L_{q_v}(\Theta'_v) = ||F(v; \Theta'_v) - h_0||^2.$$  \hspace{1cm} (9)

Subsequently, the prior $\Theta$ can be updated during meta-training by minimize the query-based total loss of all meta-training tasks. Given the set of meta-training tasks $T_{train}$, we optimize the following to obtain the optimal prior:

$$\arg\min_{\Theta} \sum_{T_v = (S_v, q_v) \in T_{train}} L_{q_v}(\Theta - a \frac{\partial L_{S_v}(\Theta)}{\partial \Theta}).$$  \hspace{1cm} (10)

On the other hand, in our meta-testing, consider a task $T_u = (S_u, q_u)$. The prior $\Theta$, learned from meta-learning, will be adapted on the support set $S_u$ to produce a local model $\Theta'_u$ in the same way as Eq. (8). The local model $\Theta'_u$ will be simply applied on the query $u$, which is a tail node, to predict a new embedding vector $h_u = F(u; \Theta'_u)$ as the output of meta-tail2vec.

**Optimization and complexity.** To adapt to the support set of each task in Eq. (8), we apply the standard gradient descent with one or a few steps. To minimize the meta-objective in Eq. (10), we adopt the widely used Adam optimizer. As training is performed over mini-batches of tasks, the time cost of our meta-training process depends on the total number of tasks $N$ passed through, where each task contains up to $k$ support nodes and the embedding of each node is aggregated from $m$-hop neighbors. Thus, the overall complexity is $O(Nkd^m)$, where $d$ is the average degree of nodes. Typically $m$ is a small constant such as 1 or 2, and $k$ is also small by the definition of tail nodes. Furthermore, it is also common to perform neighborhood sampling [12] during the $m$-hop aggregation, and thus the average degree $d$ is also effectively restricted to a constant.

## 5 EXPERIMENTS

In this section, we conduct node classification and link prediction on three public benchmark datasets, and evaluate the performance of our proposed meta-tail2vec.

### 5.1 Experimental settings

#### 5.1.1 Datasets.** We conduct experiments on three public datasets, as follows. (1) Wiki [39] is a network of Wikipedia pages, where each node is a page, and each edge represents the hyperlink between pages. Each page belongs to one of the 19 categories. (2) Flickr [26] is a network of users of the photo sharing service, where each node is a user, and each edge represents the friendship between users. Every user belongs to one or more interest groups, such as “scenic photos”. (3) Email [41] is an e-mail network between members of a European research institution, where each node is a member, and each edge represents the communication between members. Every member belongs to one of the 42 departments. Their statistics are summarized in Table 1. Note that we regard nodes with 5 or fewer links as the tail nodes, i.e., $|\mathcal{V} : |\mathcal{N}_u| \leq 5|$.**

<table>
<thead>
<tr>
<th>Dataset</th>
<th># nodes</th>
<th># edges</th>
<th># node classes</th>
<th>multi-label</th>
<th># tail nodes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wiki</td>
<td>2,405</td>
<td>17,981</td>
<td>19</td>
<td>No</td>
<td>1,069</td>
</tr>
<tr>
<td>Flickr</td>
<td>80,513</td>
<td>5,899,882</td>
<td>195</td>
<td>Yes</td>
<td>9,367</td>
</tr>
<tr>
<td>Email</td>
<td>1,005</td>
<td>25,571</td>
<td>42</td>
<td>No</td>
<td>235</td>
</tr>
</tbody>
</table>

### 5.1.2 Base embedding models.** Our approach meta-tail2vec is embedding-agnostic and can work with any embedding model. In our experiments we examine two broad categories of base embedding models. First, we consider classic network embedding and graph neural networks.

- **DeepWalk [26]:** a pioneering, widely adopted network embedding model, which samples an equal number of paths from each node, and feeds these paths into a skip-gram model to learn node embeddings.
• GraphSAGE [12]: a graph neural network that performs graph convolutions to aggregate features from neighboring nodes recursively. For node features, we use node embedding vectors from DeepWalk and node degrees. We adopt its self-supervised version to learn the initial base embeddings.

Second, we consider robust models designed for sparse networks.

• SDNE [34]: a deep network embedding model that is robust for sparse network, by incorporating global network structures in addition to local structures.
• ARGA [25]: an adversarially regularized graph autoencoder, which achieves robust embedding by learning the data distribution of latent codes on the graph.
• DDGCN [3]: a graph convolutional network with both dual dropouts at the node and edge levels, which is more effective in reducing overfitting on sparse networks.

Setup and Parameters. To ensure the base models achieve their respective optimal performance, we performed a grid search to tune their parameters (optimal values in italics). For DeepWalk, we searched the number of walks \( y \in \{5, 10, 20\} \), walk length \( r \in \{40, 100, 150\} \) and window size \( w \in \{3, 5\} \). For GraphSAGE, we adopted a two-layer architecture, chose the aggregator from \{mean, meanpool, maxpool\} and tuned the dimension of hidden layer over \{32, 64, 128\}. For SDNE, we searched the weight of local structures (as opposed to global structures) \( \alpha \in \{50, 100, 150\} \) and the weight of reconstruction \( \beta \in \{10, 30, 50\} \). For ARGA, we tuned the dimension of hidden layer over \{16, 32, 64\}. For DDGCN, we tuned the dropout probability \( p \in \{0.1, 0.3, 0.5\} \), and dual-dropout coefficient \( \alpha \in \{0.1, 1, 5\} \). The optimal parameters found are generally consistent with the recommended values in the literature. For all models, the dimension of embedding vectors is set to 128.

5.1.3 Baselines for tail node refinement. We employ a series of baselines that are also designed to improve tail node embeddings.

• Biased walk: Since tail nodes are under-represented, we oversample paths starting from the tail nodes. This method is only applicable to the base embedding models DeepWalk and GraphSAGE, since the other base models do not utilize random walks to sample node pairs.
• Additive [18]: Aggregates the embeddings of the neighboring nodes as the output embedding for a tail node. We also compare to Additive-2 which aggregates the embeddings of neighboring nodes within 2 hops.
• a la carte [16] is a further extension of the additive model, with a transformation through an auxiliary regression task. Similarly, we also compare to a a la carte-2, which aggregate the embedding of 2-hop neighbors.
• Nonce2vec [14] constructs a better initialization with the additive vectors, and performs another round of training using the corresponding base embedding model.
• Dropout: Inspired by DropoutNet [33], for each head node we only sample its \( k \) neighbors to feed into an auxiliary regression task.

Setup and Parameters. The goal is to predict new embedding vectors for the tail nodes by meta-tail2vec and each baseline refinement method, w.r.t. the initial embedding vectors from each of the base embedding models.

For biased walk, we over-sampled the paths starting from the tail nodes by doubling the number of random walks compared to head nodes. For Additive, a la carte and Nonce2vec, we used mean pooling as the aggregation function, which yields better empirical performance than min or max pooling. For a la carte and Dropout, the auxiliary regression tasks used the same regression model in our approach. For Nonce2vec, with DeepWalk, SDNE and ARGA as the base embedding model, the improved initialization was directly used as a pre-training; with GraphSAGE and DDGCN, the improved initialization was used as nodes’ initial feature vectors.

For our method meta-tail2vec, we set \( \alpha \), the local learning rate of adapting to the support set to 0.01, and set the global meta-learning rate to 0.001. Note that it is typical to employ a larger local learning rate in order to achieve stable training [9, 43]. The number of gradient updates during adaptation was set to 5, noting that few updates such as 1 or 3 give very close results. In the regression model \( F \), we set the size of the hidden layer of the MLP to 1024, and aggregated nodes within 2 hops. We will also analyze the impact of the number of hops in Sect. 5.3.

5.1.4 Downstream applications. We experimented with two downstream applications on the three datasets.

Node Classification. We carry out multi-class classification on Wiki and Email where each node belongs to exactly one class, and multi-label node classification on Flickr where each node can belong to one or more classes. Specifically, we evaluated the classification performance on the tail nodes, after training a logistic regression classifier on the head nodes. The predicted embeddings of the tail nodes and the oracle embeddings of the head nodes were used as their features, respectively. We adopted the evaluation metrics of micro-F and accuracy.

Link Prediction. For nodes with 2–6 links, we adopted the common leave-one-out strategy by first removing a random link from each of them (which becomes a tail node with 5 or fewer links). Our goal is to predict the removed link. On the partial network, we constructed initial embedding vectors using each base embedding model, and predicted new embedding vectors for the tail nodes with our method and each baseline refinement method. For each tail node, we treated the removed link as a positive candidate, and randomly sampled five other non-linked nodes as negative candidates. Link prediction was then formulated as a ranking problem: given a tail node, we rank its candidates using a learning-to-rank model [8] trained on the head nodes. Likewise, the predicted embeddings of the tail nodes and the oracle embeddings of the head nodes were used as their features, respectively. We adopted the evaluation metrics of mean reciprocal rank (MRR) and hit ratio at top 1 (Hit@1).

5.2 Performance comparison

We present the performance of our meta-tail2vec and various baseline refinement methods, w.r.t. each base embedding model. As the primary goal of this paper is to improve tail node embeddings, we mainly focus on comparing the performance on the tail nodes. Nevertheless, to further demonstrate that head node embeddings are not adversely impacted, we also investigate the performance on the head
nodes. Note that all results are averaged over 10 runs and reported with their standard deviations; best method appears in bold, and the second best approach is underlined.

### 5.2.1 Comparison of tail node embeddings.

We first study classic base embedding models that are not specifically designed for robustness on sparse networks, followed by robust base models designed for sparse networks.

**Classic base models.** We report the performance of node classification in Table 2 w.r.t. classic base embedding models DeepWalk and GraphSAGE, respectively. On the one hand, our meta-tail2vec achieves significant improvements over the base embedding methods consistently, by 7.7%–10.9% w.r.t. DeepWalk and 9.3%–13.3% w.r.t. GraphSAGE in terms of MicroF. The results demonstrate that tail node embedding is a critical problem to address, and our proposed approach is indeed useful in refining the tail node embeddings. On the other hand, meta-tail2vec also outperforms other refinement approaches w.r.t. each of the base models SDNE, ARGA and DDGCN. While refined models achieve significant improvements over the base embedding methods, our model also aggregates 2-hop nodes and attains better performance than its 1-hop version (as we will see in Sect. 5.3), potentially due to the local adaptation which can effectively filter noises at each locality.

We further report the performance of link prediction in Table 3 w.r.t. classic base embedding models. Similar conclusions can be drawn, where meta-tail2vec outperforms the base models by 3.1%–11.9% in MRR and 6.2%–24.9% in MicroF, respectively.

**Robust base models for sparse networks.** We also investigate whether meta-tail2vec can also improve base models designed for robustness on sparse networks.

We report the performance of node classification in Table 4, w.r.t. each of the base models SDNE, ARGA and DDGCN. While these base models are intended to handle sparse networks, they aim to increase the overall robustness of the learning process, and do not explicitly improve the embedding of the most vulnerable tail nodes. Thus, their performances on the tail nodes are not necessarily better.
than classic base models. Note that our approach meta-tail2vec is embedding-agnostic, meaning that even for base models already designed for sparse networks, we can still refine their tail node embeddings, as demonstrated by the results that meta-tail2vec outperforms the base embeddings by an average of 14.9% in terms of MicroF on node classification. On the other hand, we also compare meta-tail2vec to other baseline refinement methods. (Due to space constraint, we only present the results of Addictive, Nonce2vec and Dropout, which are generally the best baselines among all.) Again, due to our locality-aware task formulation, the meta-learning strategy is able to adapt to the locality of each tail node well, resulting in an average performance lift of 8.5% in terms of MicroF compared to the best baseline.

Furthermore, we report the performance of link prediction in Table 5. We observe similar performance comparisons, where on average meta-tail2vec outperforms the robust base models by 5.2% and the best baseline by 2.0% in terms of MRR.

<table>
<thead>
<tr>
<th>Table 4: Performance of node classification w.r.t. robust base embedding models (MiF for MicroF; Acc for accuracy).</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base</td>
</tr>
<tr>
<td>Wiki</td>
</tr>
<tr>
<td>Acc</td>
</tr>
<tr>
<td>Flickr</td>
</tr>
<tr>
<td>Acc</td>
</tr>
<tr>
<td>Email</td>
</tr>
<tr>
<td>Acc</td>
</tr>
<tr>
<td>DDGCN as the base embedding model</td>
</tr>
<tr>
<td>Acc</td>
</tr>
<tr>
<td>Flickr</td>
</tr>
<tr>
<td>Acc</td>
</tr>
<tr>
<td>Email</td>
</tr>
<tr>
<td>Acc</td>
</tr>
<tr>
<td>ARGA as the base embedding model</td>
</tr>
<tr>
<td>Acc</td>
</tr>
<tr>
<td>Flickr</td>
</tr>
<tr>
<td>Acc</td>
</tr>
<tr>
<td>Email</td>
</tr>
<tr>
<td>Acc</td>
</tr>
<tr>
<td>MRR</td>
</tr>
<tr>
<td>Acc</td>
</tr>
<tr>
<td>Flickr</td>
</tr>
<tr>
<td>Acc</td>
</tr>
<tr>
<td>Email</td>
</tr>
<tr>
<td>Acc</td>
</tr>
</tbody>
</table>

5.2.2 Comparison of head node embeddings. While our main goal is to improve tail node embeddings, we further evaluate the performance on the head nodes to validate that their embeddings still remain competitive. In theory head node embeddings are not changed in any way, since we only predict new embedding vectors for tail nodes. However, the performance of head nodes on a downstream application can be potentially improved, as training the downstream model can still benefit from the improved quality of tail nodes.

Thus, we further conducted an experiment on head node embeddings. We sampled and evaluated a test set comprising 10% of the other nodes, inclusive the tail nodes and the remaining 90% head nodes. We tabulate the results in Figs. 4 and 5 for node classification and link prediction, respectively. As hypothesized, both meta-tail2vec and other baseline refinement approaches can slightly outperform the head node performance, due to the improved quality of tail node embeddings in the downstream training data. However, it is not surprising that the improvements are modest compared to

![Figure 4: Performance of node classification on head nodes w.r.t. DeepWalk as the base embedding model.](image-url)

![Figure 5: Performance of link prediction on head nodes w.r.t. DeepWalk as the base embedding model.](image-url)
tail nodes, in the range of 0.3%–0.9% only. The reason is that the 
head node embeddings remain unchanged and their performance is 
only indirectly influenced by refined tail node embeddings. Never-
theless, we validated the goal of significantly improving tail node 
embeddings whilst head node embeddings remain robust.

5.3 Model analysis and discussion

5.3.1 Ablation study. We analyze the contribution of our ap-
proach by an ablation study. Using DeepWalk as the base embedding 
model, we compare the following four variants of meta-tail2vec: (1) 
full, the full meta-tail2vec model; (2) global, only train one global 
embedding regression model on the head nodes, and predict the em-
bedding vectors of all tail nodes with the same global model (equiv-
alent to the dropout baseline); (3) fine-tune, fine-tune the pre-trained 
global model on the support set of a tail node before predicting 
its embedding vector; (4) rand-supp, the same approach as the full 
model except the locality awareness, which samples random nodes 
from the graph as the support sets.

Their performances are reported in Fig. 6. Among the four vari-
ants, we observe that the fine-tune model is only able to achieve 
marginal better performance than the global model, as fine-tuning 
on the small support set of a tail node can easily cause overfitting. In 
picular, on the Flickr dataset, the fine-tune model is in fact slightly 
 worse than the global model due to overfitting. Next, the rand-supp 
model performs the worst, implying that the locality-aware task 
generation is critical for graph data. Finally, the full model per-
forms the best, demonstrating the effectiveness of our locality-aware 
meta-learning approach.

5.3.2 Visualization. We study how exactly meta-tail2vec updates 
the base embeddings for the tail nodes. In particular, we visualize 
the base and refined embeddings using the t-SNE algorithm.

We first showcase two base embedding models DeepWalk and 
SDNE in Figs. 7(a) and 7(c), respectively. While both base models 
can separate the head nodes (denoted by hollow dots) into relatively 
well defined clusters corresponding to their ground truth classes (as 
coded by different colors), they fail at the tail nodes (denoted by 
solid dots). There is a significant mixture of tail nodes from different 
classes at the center of the visualization for both DeepWalk and 
SDNE, implying that they are not designed to work well on tail nodes.

On the other hand, we examine the refined embeddings by meta-
tail2vec w.r.t. the two base models in Figs. 7(b) and 7(d) respectively. 
As head nodes already have high-quality base embeddings, the key 
is to improve the tail node embeddings. Our approach specifically 
refines the tail nodes so that they now move towards the cluster 
center of their class together with the head nodes from the same 
class. The contrast with the base embeddings conclude that our 
proposed approach is effective in learning tail node embeddings.

5.3.3 Impact of tail node sparsity. We breakdown the perfor-
mance of tail nodes by their number of links. As shown in Fig. 8, the 
performance improvement is observed across the spectrum for tail 
nodes with between 1 and 5 links. Note that on the Flickr dataset, 
all nodes have at least 2 links. The improvement is generally smaller 
on nodes with only one link as expected, given that meta-tail2vec is 
also constrained by the very limited structural information.

5.3.4 Impact of neighborhood hops. Next, we analyze the im-
 pact of number of hops used for computing the input feature vector 
in Eq. (5). The results across different number of hops are presented
in Fig. 9. In particular, aggregating from the 2-hop neighborhood achieves optimal performance on all datasets. In particular, aggregating from the 3-hop neighborhood results in decreased performance due to more noises from the less relevant nodes.

6 CONCLUSION

In this paper, we investigate the problem of tail node embedding on graph, and proposed a novel approach meta-tail2vec for refining tail node embeddings. In particular, we proposed a novel formulation by casting it into a few-shot regression setting. Moreover, we designed a locality-aware task generation strategy to capture the prior knowledge across nodes at different localities. Finally, extensive experiments demonstrated the promising performance of meta-tail2vec on both node classification and link prediction.

ACKNOWLEDGEMENT

This research is supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG-RP-2018-001). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore.

REFERENCES