Inductive Link Prediction in Static and Temporal Graphs for Isolated Nodes

Ayan Chatterjee
Network Science Institute
Northeastern University
Boston, USA

Robin Walters
Khoury College of Computer Sciences
Northeastern University
Boston, USA

Giulia Menichetti
Brigham and Women's Hospital
Harvard Medical School
Boston, USA

Tina Eliassi-Rad
Network Science Institute
Khoury College of Computer Sciences
Northeastern University
Boston, USA

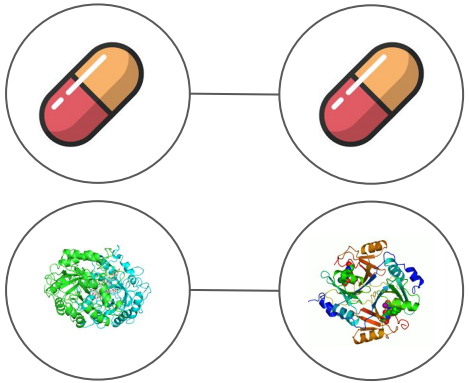
Link Prediction

- Link prediction: Anticipating connections between entities within a network.

- Applications:
  1. Drug-target interactions
  2. Protein-protein interactions
  3. Collaboration networks
  4. Citation networks
  5. Knowledge graph completion
  6. Recommender systems
  7. Dynamic routing in transportation & optical networks
  8. Intrusion detection on the internet
  9. Molecular dynamic simulations

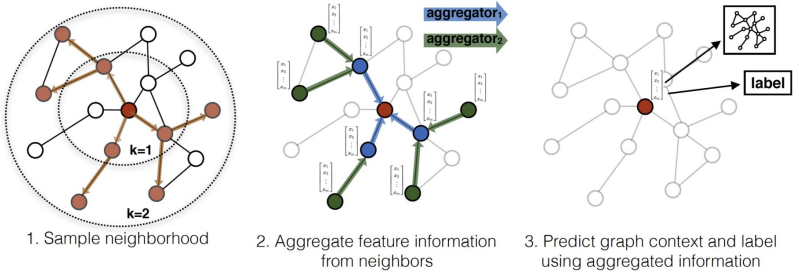
Static Graphs

Temporal Graphs



State-of-the-art in Link Prediction

- State-of-the-art static link prediction models
  - GraphSAGE, GraIL, etc.
  - Node embeddings from their neighborhood topology via aggregation.
  - Downstream decoder
  - End-to-end training



1. Sample neighborhood

2. Aggregate feature information from neighbors

3. Predict graph context and label using aggregated information

- State-of-the-art static temporal link prediction models
  - TGN, DyREP, TGAT, etc.
  - Neighborhood aggregation + memory modules (learn the temporal accumulation of neighbors)
  - Downstream decoder
  - End-to-end training

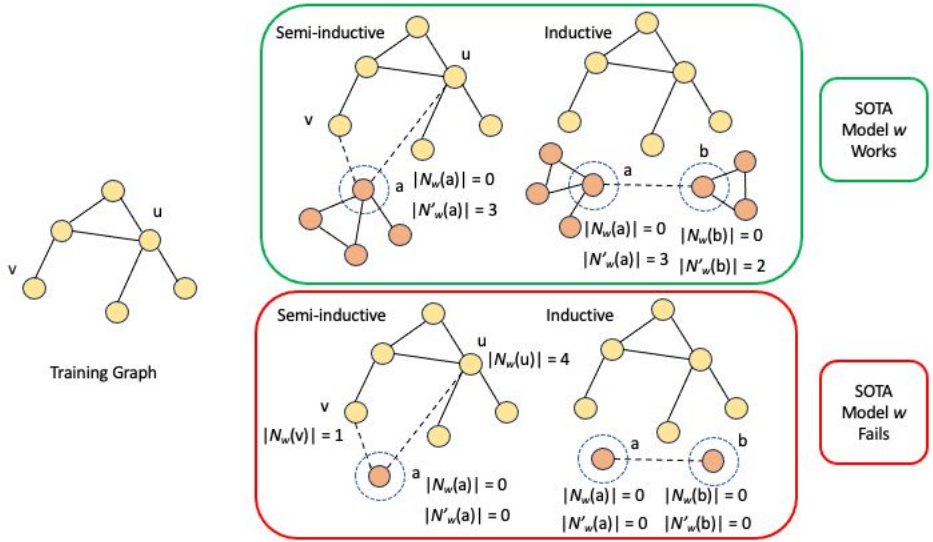
Observation Bias

- Neighborhood of a node implied available training data.
- SOTA models excel in learning nodes with abundant training data.
- These models struggle on unobserved isolated nodes that lack neighborhood topology information.
- Applications:
  - Studying binding between a new drug and a novel target
  - Exploring interactions between poorly-annotated proteins

  Exploring rare diseases
  - Recommending a new product
  - Introducing a new user in a social network
  - Introducing a new wireless device in a wireless network

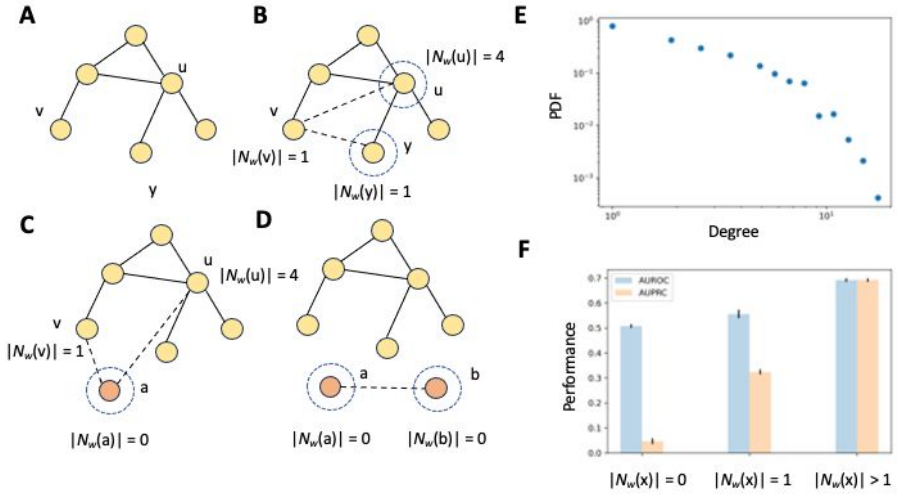
    … …

Observation Bias



$N_w(a)$ and $N'_w(a)$ represent the neighborhoods of node a observed by the link prediction model w during train and test, respectively. We are interested in two link prediction scenarios: semi-inductive (when one node of the test edge is unobserved in training) and inductive (when both nodes of the test edge is unobserved in training). SOTA models prove ineffective for nodes that were not observed during training and possess inadequate neighborhood information during testing.

GraphSAGE on Isolated Nodes



GraphSAGE on benchmark OGB drug-drug interaction network dataset. The drug-drug interaction network showcases a power-law degree distribution. Therefore, the majority of the nodes in the network have low-degrees, which coexist with some high-degree nodes or hubs. GraphSAGE achieves excellent link prediction performance for the nodes with high degrees in the DDI graph. For the nodes with only one observed edge in training, the performance drops significantly. GraphSAGE fails in making correct link predictions for the nodes with no neighborhood data.

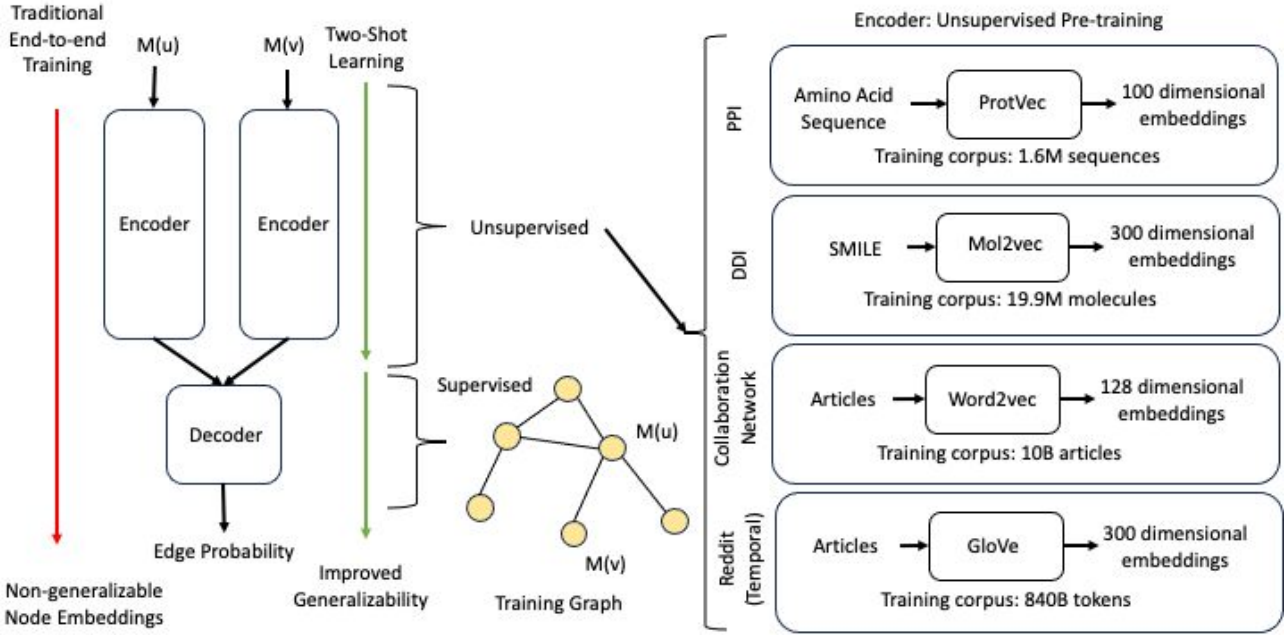
Our Method

- We propose a two-shot learning approach involving unsupervised pre-training of node attributes on a corpus different from and larger than the observed graph that improves inductive link prediction performance on isolated nodes.

- Instead of training the link prediction model in an end-to-end fashion (node embeddings and downstream decoder altogether) we first train the node features in an unsupervised fashion on entities different from the nodes of the training graph, and thereafter use the train graph to learn a downstream link prediction decoder.

- We obtain a generalizable representation, independent of the train graph topology, suitable for never-before-seen isolated nodes in test.

Our Method



OPEN GRAPH BENCHMARK

Graph	Number of Nodes
PPI	576,289 proteins
DDI	4,267 drugs
Collab	300M papers, 235,868 researchers
Reddit	3.6B tokens, 118,381 users, 51,278 subreddits

Results: Static and Temporal Inductive Link Prediction

PLNLP is a SOTA static link prediction model from OGB leaderboard. Our approach significantly improves inductive link prediction performance on unseen nodes.

Dataset	PLNLP			Pre-trained Node Attributes		
	AUROC	AUPRC	Hits@TopK(%)	AUROC	AUPRC	Hits@TopK(%)
ogbl-ppa	0.51 ± 0.03	0.12 ± 0.04	0.09 ± 0.03	0.78 ± 0.03	0.35 ± 0.03	0.39 ± 0.03
ogbl-collab	0.61 ± 0.03	0.23 ± 0.07	11.56 ± 0.93	0.97 ± 0.02	0.92 ± 0.02	36.44 ± 3.11
ogbl-ddi	0.50 ± 0.04	0.11 ± 0.07	0.01 ± 0.02	0.54 ± 0.02	0.21 ± 0.02	0.39 ± 0.02

DyHATR is a temporal link prediction model combining node representation learning with memory modules. Our approach improves temporal inductive link prediction performance on unseen nodes significantly.

Model	2014-2015		2016-2017		2017-2018	
	AUROC	AUPRC	AUROC	AUPRC	AUROC	AUPRC
Pre-trained Node Attributes	0.70	0.66	0.63	0.60	0.69	0.65
DyHATR	0.45	0.25	0.45	0.48	0.46	0.28

Takeaways

- SOTA link prediction models leverage neighborhood topology
- These models are trained end-to-end
- Neighborhood of a node represents data availability
- SOTA models fail for unobserved isolated nodes in inductive link prediction
- We propose non-end-to-end training for never-before-seen nodes
- We train node attributes on a corpus larger than the training graph in an unsupervised manner