Shopping and Discovery

Inductive Graph Neural Networks for Transfer Learning

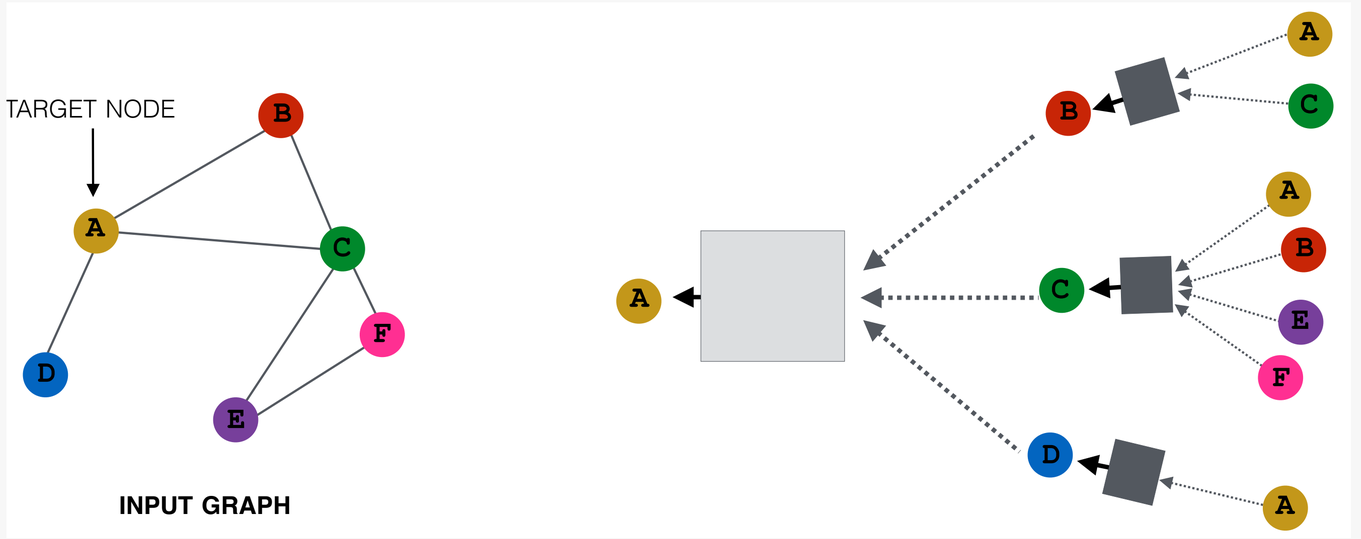
Atlas Wang, UT Austin & Amazon

Together with: **Wenqing Zheng, Yan Han, Nikhil Rao, Eddie Huang, Karthik Subbian**

Many slides credits from Karthik

Graph Convolutional Networks (GCN)

- Aggregate messages from the neighbors
- Apply non-linear activation



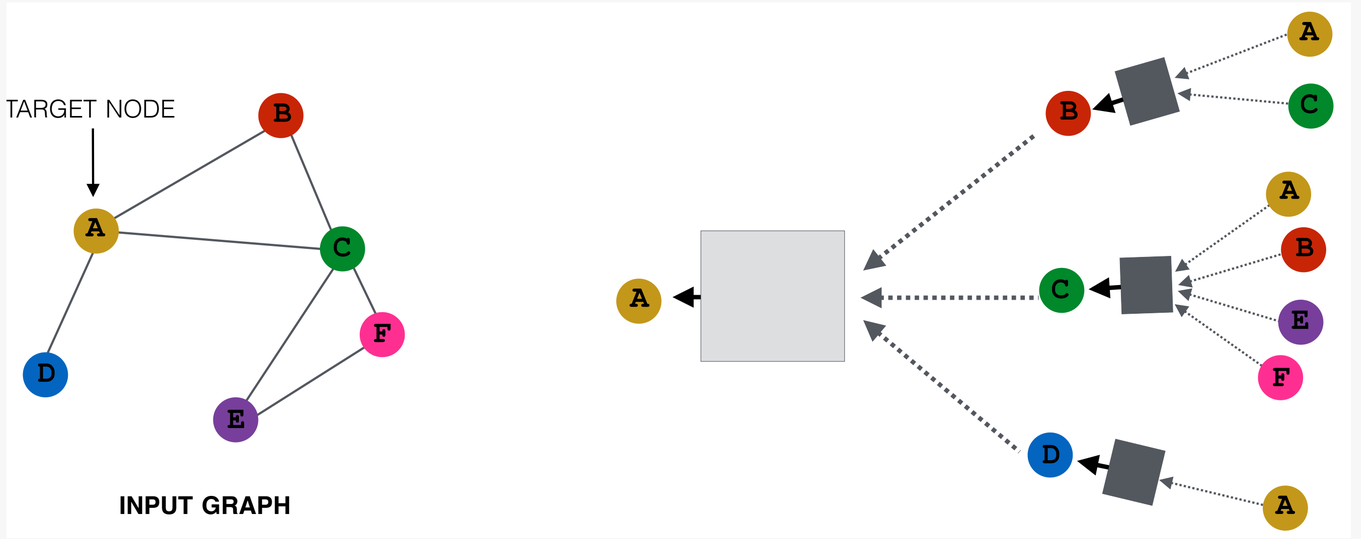
$$\mathbf{h}_v^0 = \mathbf{x}_v$$

$$\mathbf{h}_v^k = \sigma\left(\mathbf{W}_k \sum_{u \in N(v)} \frac{\mathbf{h}_u^{k-1}}{|N(v)|} + \mathbf{B}_k \mathbf{h}_v^{k-1}\right), \quad \forall k > 0$$

Scarselli et al. 2005. The Graph Neural Network Model. *IEEE Transactions on Neural Networks*.

Graph Convolutional Networks (GCN)

- Aggregate messages from the neighbors
- Apply non-linear activation



$$\mathbf{h}_v^0 = \mathbf{x}_v$$

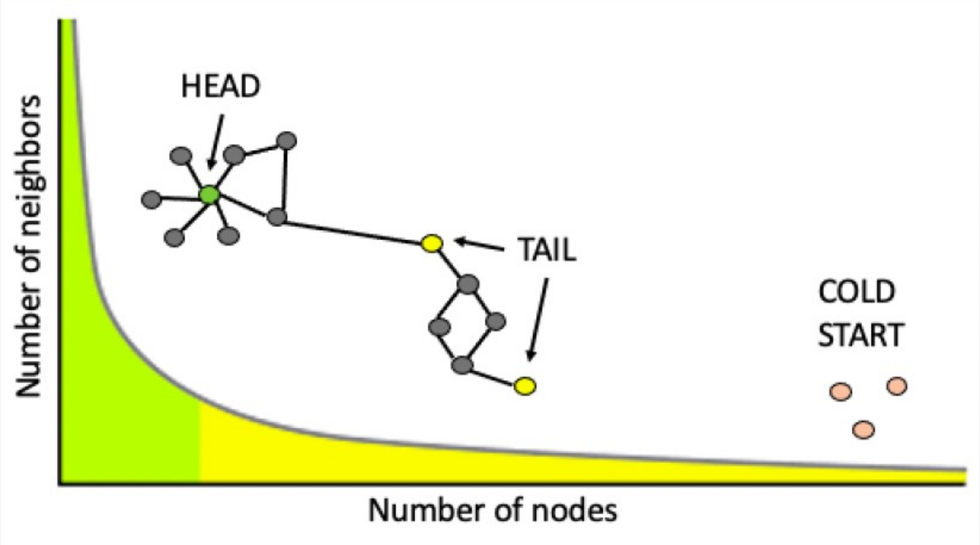
$$\mathbf{h}_v^k = \sigma\left(\mathbf{W}_k \sum_{u \in N(v)} \frac{\mathbf{h}_u^{k-1}}{|N(v)|} + \mathbf{B}_k \mathbf{h}_v^{k-1}\right), \quad \forall k > 0$$

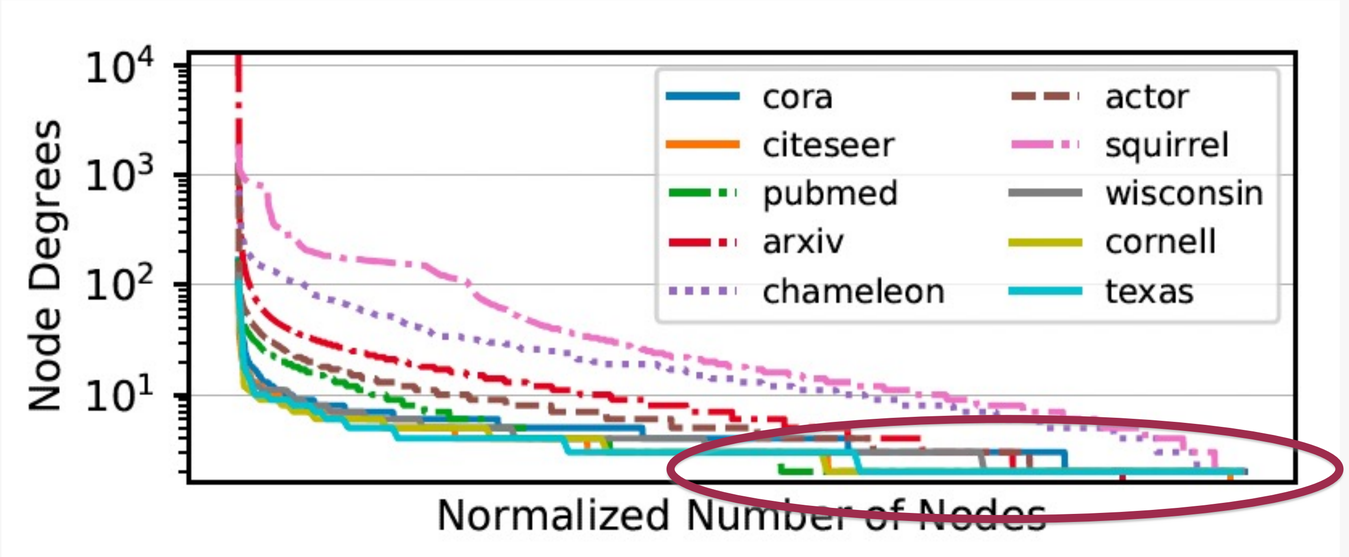
Scarselli et al. 2005. The Graph Neural Network Model. *IEEE Transactions on Neural Networks*.

Cold-Start Problem in Search

- Tons of new products (ASINs) added each day to our Catalog
- These ASINs often lack detailed product information (e.g., product type) and behavioral signals
- Emerging and new locales have high number of cold start ASINs
- Goal: achieve 100% product type signal coverage

Cold-start Problem in Graphs





- Cold start nodes are common in natural graphs
- They do not have enough neighbors to make use of existing GNN models
- Neighbors are noisy or non-existent

Existing Inductive GNN Approaches

Inductive GNN approaches need the knowledge of the network and assume robust edges

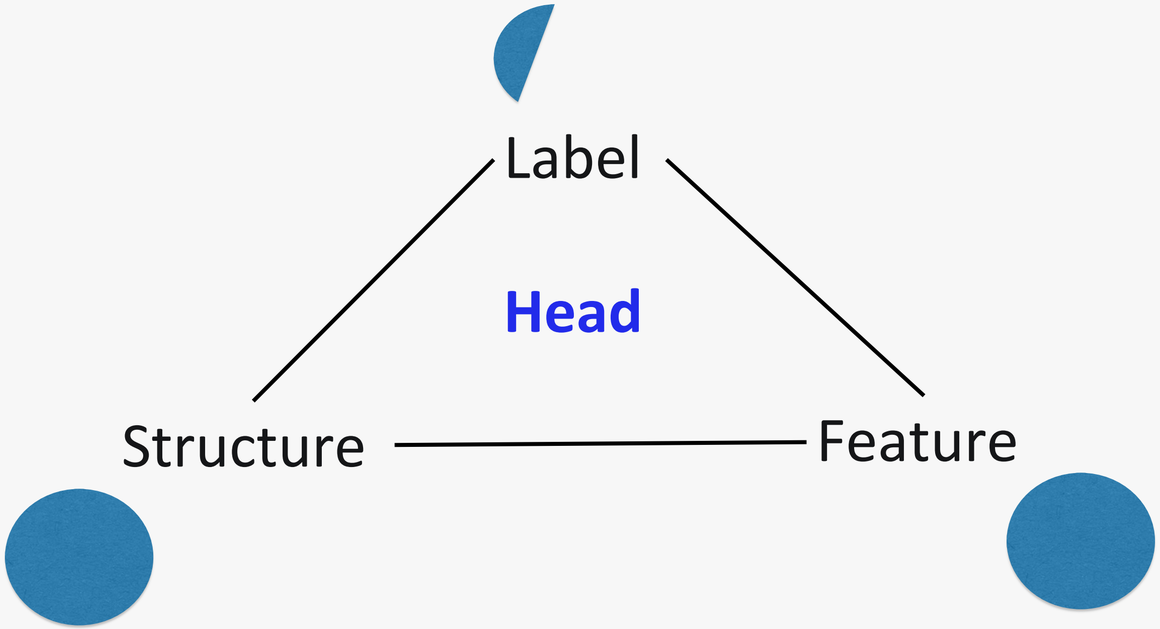
e.g., GraphSage, Hamilton et. al, 2017.

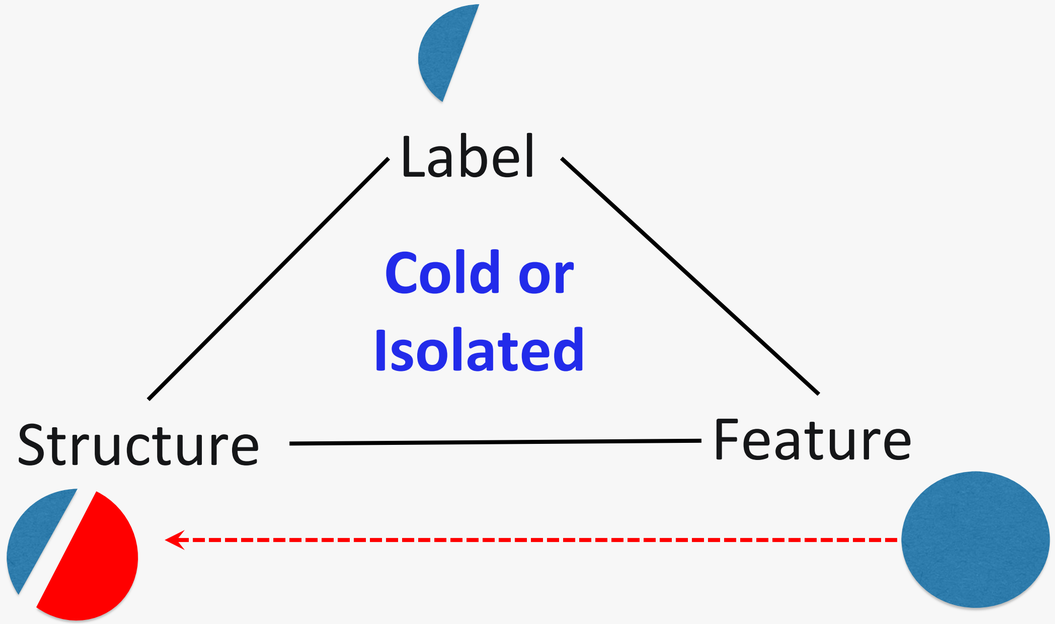
Models that use the graph structure without message passing to learn the representation

e.g., GraphMLP, Hu et. al, 2021.

None of these techniques work well for cold-start or isolated nodes in the graph

Head vs cold-start Nodes





Learning efficient mapping from features to fill the structural gap

Feature and Label Smoothing Theorem*

$$\mathbf{x}_i = \frac{1}{d_{ii}} \sum_{j \in \mathcal{N}(i)} a_{ij} \mathbf{x}_j + \epsilon_i$$

Error from feature-structure correlation

Large error for cold start nodes

$$\left| y_i - \frac{1}{d_{ii}} \sum_{j \in \mathcal{N}(i)} a_{ij} y_j \right| \leq L \|\epsilon_i\|_2 + o\left(\max_{j \in \mathcal{N}(i)} (\|\mathbf{x}_j - \mathbf{x}_i\|_2) \right)$$

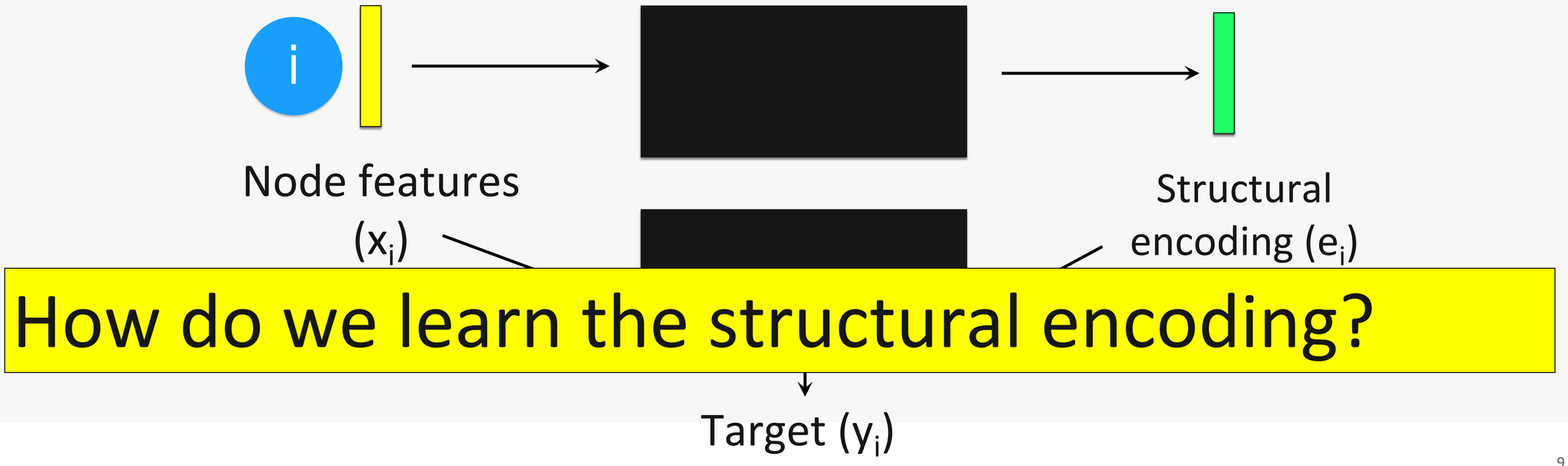
Labels of node i and their neighbors are uncorrelated when error is large

High quality structural information is critical for cold start nodes

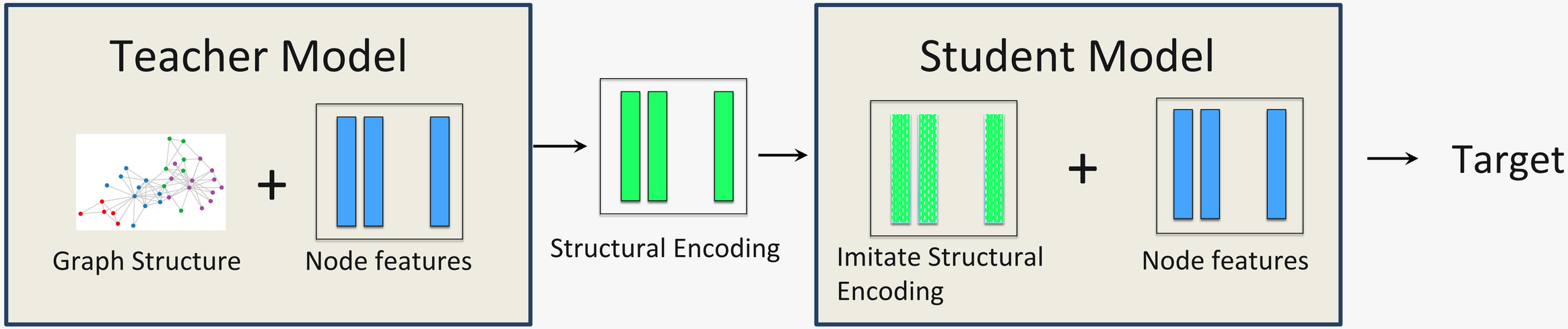
H. Wang, J. Leskovec: **Unifying Graph Convolutional Neural Networks and Label Propagation. ArXiv, 2020.*

Our Idea

1. Learn to map the node features of node i to structural encoding
2. Structural encoding should not depend on the neighborhood of i
3. Then structural encoding along with node features can be used to predict the target



Our Approach: Cold-Brew Overview



- Our approach is a student-teacher distillation model
- Teacher extracts the structural encoding and student imitates it
- Student uses the imitated structural encoding with features to predict target
- This is different from standard student models which mimic target

Problem Setting

- Notations
  - Adjacency Matrix $A \in \Re^{N \times \mathrm{N}}$
  - Degree Matrix $D \in \Re^{N \times \mathrm{N}}, where\ d_{ii} \geq 0, d_{ij} = 0$
  - Normalized adjacency matrix $\tilde{A} = D^{-1/2} A D^{-1/2} \in \Re^{N \times \mathrm{N}}$
  - Learned representations at layer ℓ $X^{(l)} \in \mathbb{R}^{N \times d_1}$
  - Neighborhood of node i N_i
  - Weight Matrix learned at layer l $W^{(l)} \in \mathbb{R}^{d_1 \times d_2}$

- Inductive node classification problem
  - Given: Node labels and graph structure
  - Goal: Infer accurate labels for cold start and isolated nodes

Teacher Model

$$\mathbf{x}_i = \frac{1}{d_{ii}} \sum_{j \in \mathcal{N}(i)} a_{ij}\mathbf{x}_j + \epsilon_i$$

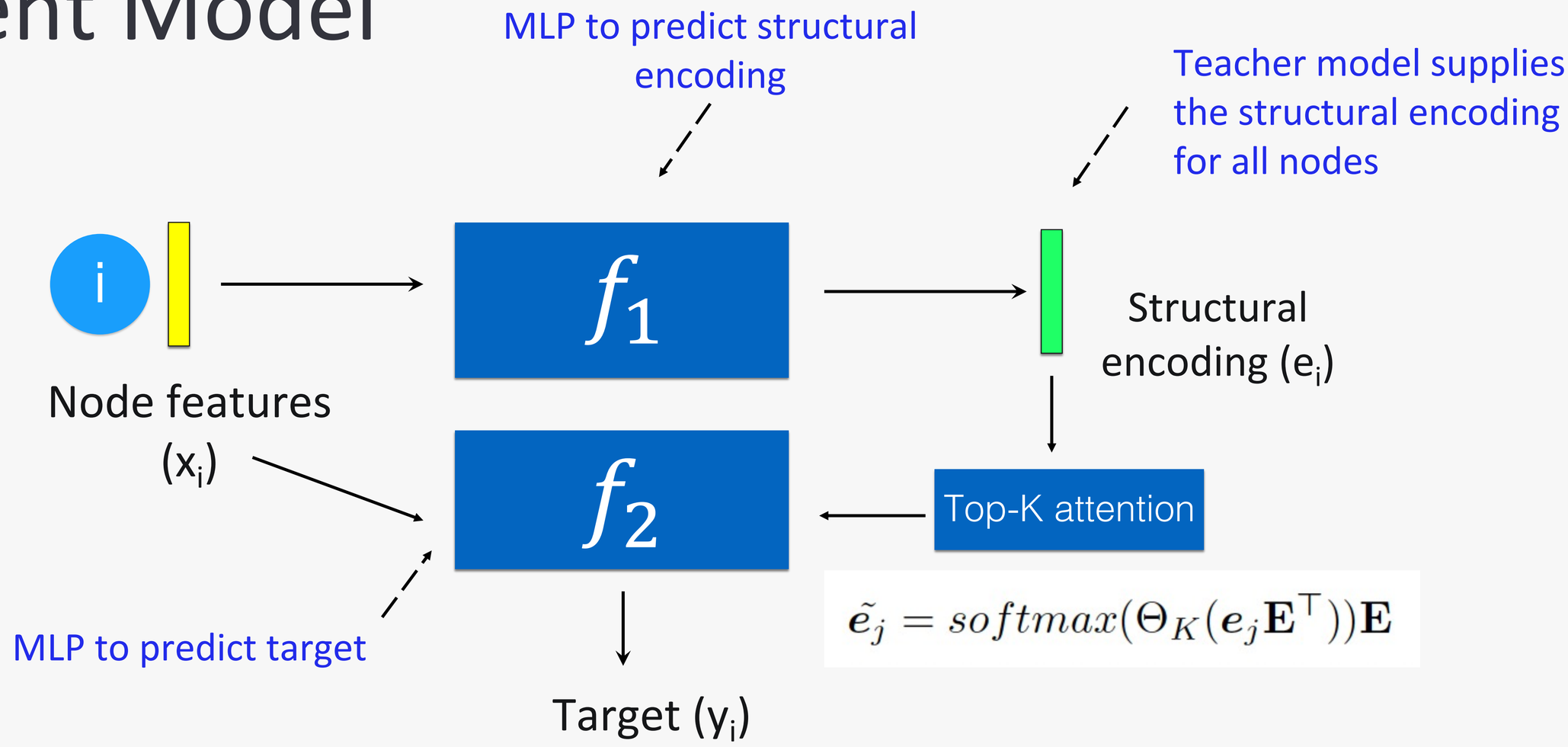
SE-GCN model $$X^{(l+1)} = \sigma(\tilde{A}(X^{(l)}W^{(l)} + E^{(l)}))$$

Structural encoding = Feature encoding + Label encoding

Loss function $$CE\left(X_{train}^{(L)}, Y_{train}\right)$$

Label encoding helps improves label smoothing

Student Model



Feature Contribution Ratio (FCR)

- We introduce FCR measure to understand the contribution of node features to the GNN model compared to the edge features

$$\delta_{MLP} = z_{GNN} - z_{MLP}, \quad \delta_{LP} = z_{GNN} - z_{LP}$$

$$FCR(\mathcal{G}) = \begin{cases} \frac{\delta_{LP}}{\delta_{MLP}+\delta_{LP}} \times 100\% & z_{MLP} \leq z_{GNN} \\ 1 + \frac{|\delta_{MLP}|}{|\delta_{MLP}|+\delta_{LP}} \times 100\% & z_{MLP} > z_{GNN} \end{cases}$$

- FCR = 0 implies graph structure contributes entirely to the model performance
- FCR = 100 means node feature contributes entirely to the model performance

Experiments: Baselines

- **GNNs**
  - GCN 2-layer
- **MLPs**
  - SimpleMLP
  - GraphMLP
- **ColdBrew**
  - Teacher (GCN + SE 2-layers)
  - Student (MLP)

More experiment and dataset details can be found in our paper: "*Cold Brew: Distilling Graph Node Representations with Incomplete or Missing Neighborhoods*", **ICLR 2022**

Experiments: Isolated Nodes

Method	Ecomm1	Ecomm2	Ecomm3	Ecomm4
GCN 2 layers	0	0	0	0
Simple MLP	+5.89	+9.85	+5.83	+6.42
Graph MLP	+6.27	+9.46	**+5.99**	+7.37
Teacher (GCN + SE 2 layers)	+0.27	+0.76	-0.50	+1.22
Student (MLP)	**+7.56**	**+11.09**	+5.64	**+9.05**

- Student MLP improves accuracy on isolated nodes up to +11%

Experiments: Tail (or cold start) Nodes

Method	Ecomm1	Ecomm2	Ecomm3	Ecomm4
GCN 2 layers	0	0	0	0
Simple MLP	-0.37	+1.74	-0.13	-0.45
Graph MLP	-0.33	+1.64	**+1.27**	+0.80
Teacher (GCN + SE 2 layers)	**+0.85**	+0.44	-0.60	+1.10
Student (MLP)	+0.32	**+3.09**	-0.18	**+2.09**

- Student model outperforms baselines by +3%
- Teacher model begins to do well compared to student model with more graph information

Experiment: Isolated Vs Tail Vs Head

Method	Isolated	Tail	Head
GCN 2 layers	0	0	0
Simple MLP	+5.89	-0.37	-16.74
Teacher (GCN + SE 2 layers)	+0.27	+0.85	+1.11
Student (MLP)	+7.56	+0.32	-15.28

- Teacher model consistently outperforms GCN and increasingly better
- Student model really well for isolate and tail nodes
- As we get more structural information GCN performs improvingly better and significantly outperform MLP models

Conclusion

- We proposed a knowledge distillation framework to extract useful structural embeddings for cold start problem
- The teacher model extracts the structural embeddings
- The student model mimics the embeddings purely from the node features
- Our approach is robust to deep GCN over smoothing problems
- Our approach is scalable and well suited for cold-start and isolated nodes