Featureless Graph Data Predicting

Speaker: Ruiwen Zhou (Group 12)

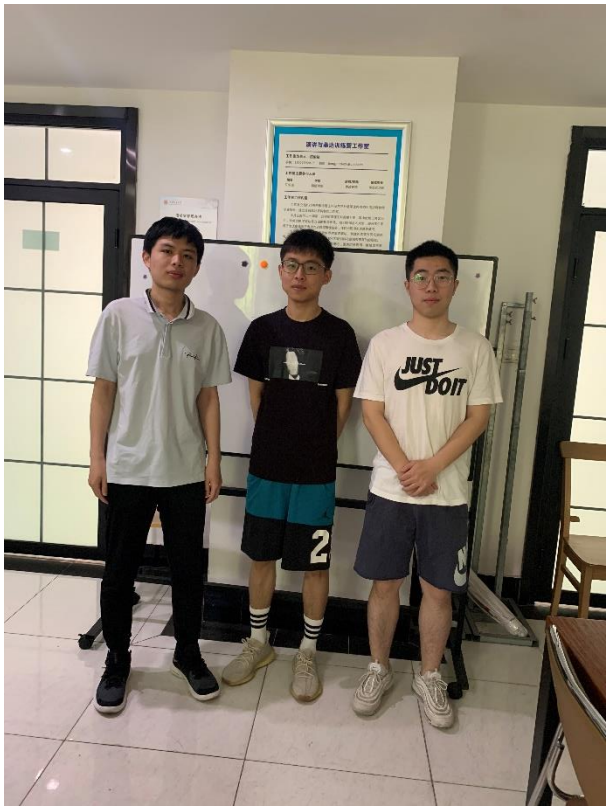
June, 2021

上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

Group Division

- From the left to the right:

Group Member	Major Contribution
Ruiwen Zhou	Literature Research Link Prediction
Rui Ye	Data Preprocessing Node Classification
Zhiyu Zhang	Ensemble Learning Scheme Node Classification

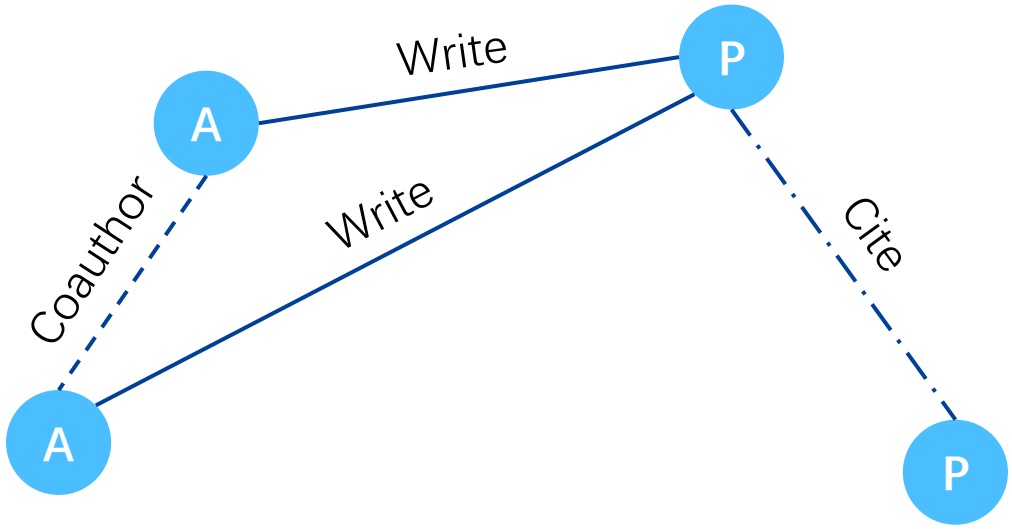


- Although in general we work in a parallel manner, we communicate and discuss on both directions often.

Contents

AceMap Network Modelling

- We build a **HOMOGENEOUS** network
- Involving **nodes of both types**: Papers and Authors



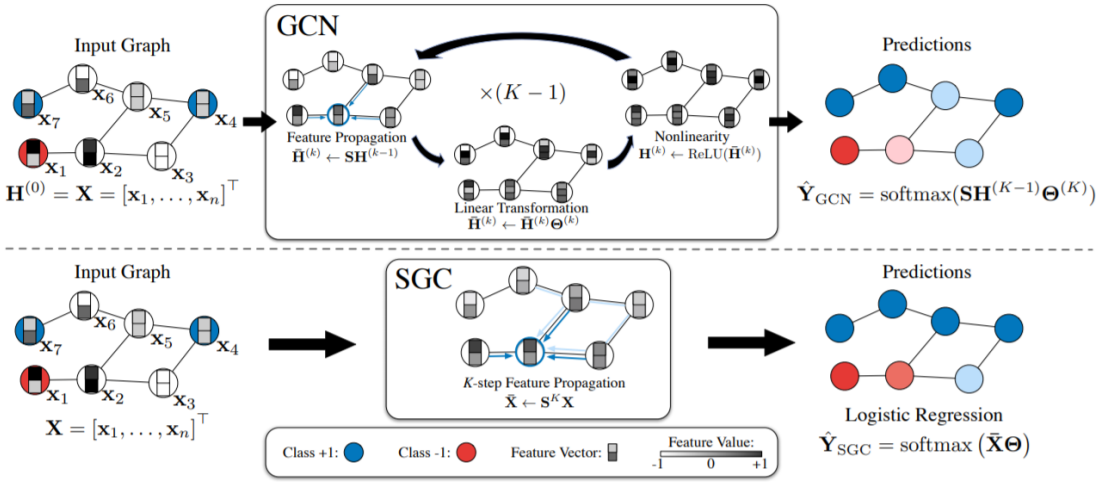
- **Three edge types**: Citation, Authorship, and Co-author relationship

Node Classification Configuration

Feature Engineering

Labelled paper	0	0	0	0	0	1	0	0	0	0	1	0	0
Unlabeled paper	0	0	0	0	0	0	0	0	0	0	0	1	0
Author	0	0	0	0	0	0	0	0	0	0	0	0	1

GCN & SGC



F. Wu, T. Zhang, et al. Simplifying Graph Convolutional Networks. ICML 2019.

SEAL Framework

- We follow the **SEAL** framework proposed in NeurIPS 2018

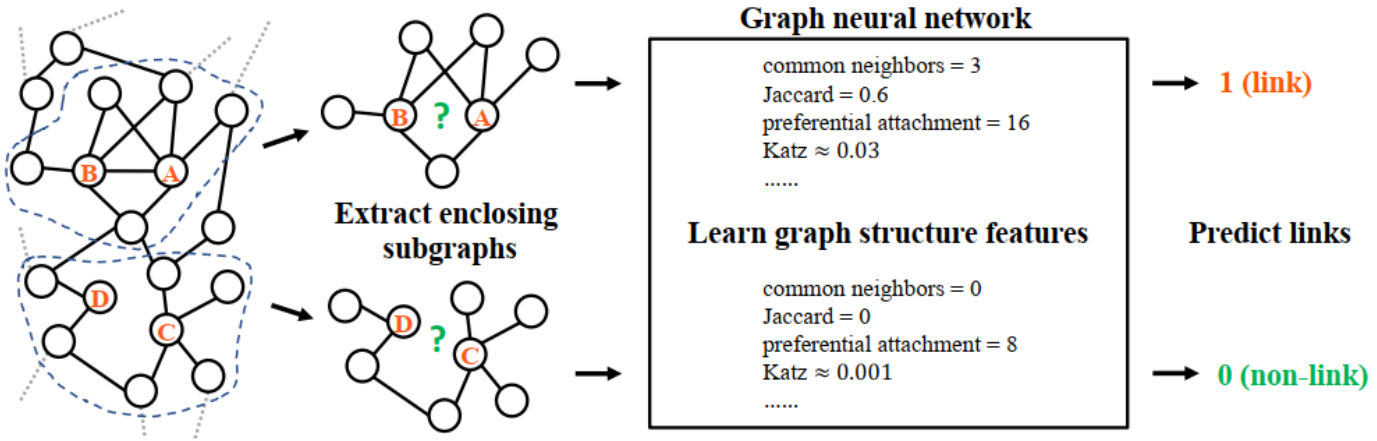


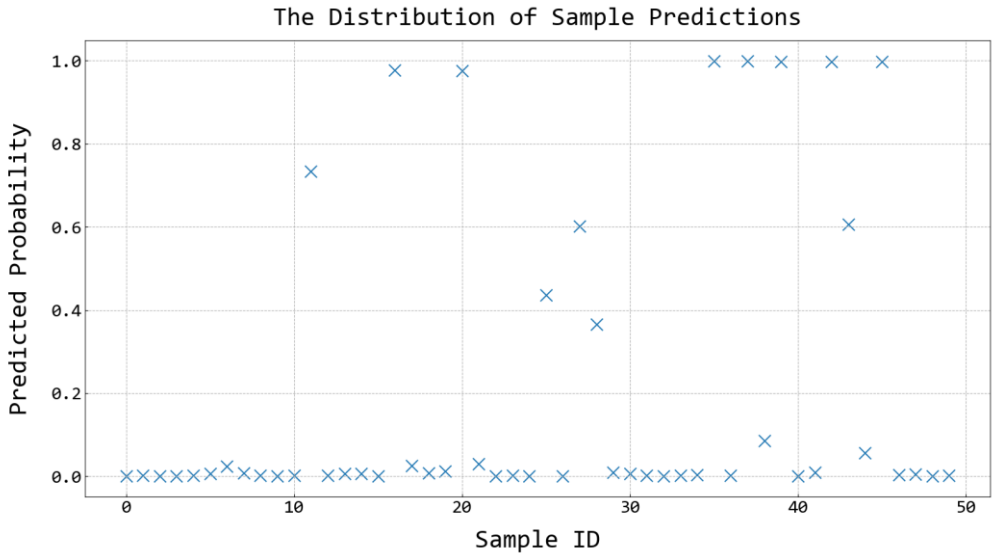
Figure 1: The SEAL framework. For each target link, SEAL extracts a local enclosing subgraph around it, and uses a GNN to learn general graph structure features for link prediction. Note that the heuristics listed inside the box are just for illustration – the learned features may be completely different from existing heuristics.

- And we replace the DGCNN in SEAL by **Hierarchical ASAP** Pooling Net

M. Zhang, and Y. Chen. Link Prediction Based on Graph Neural Networks. NeurIPS 2018.

Overfitting Problem

- Using unweighted edges and DGCNN (as default setting in SEAL)
- We obtain some prediction distributed like this



- Most prediction falls into a narrow range centering at 0 and 1
- Poor generalization, which does great harm to AUC score

Trick 1: Soft Labels

- Assume a pair of authors (a_i, a_j) coauthor n_{ij} papers
- Requirement:

More cooperation → More Determined Label

- The label of this author pair is

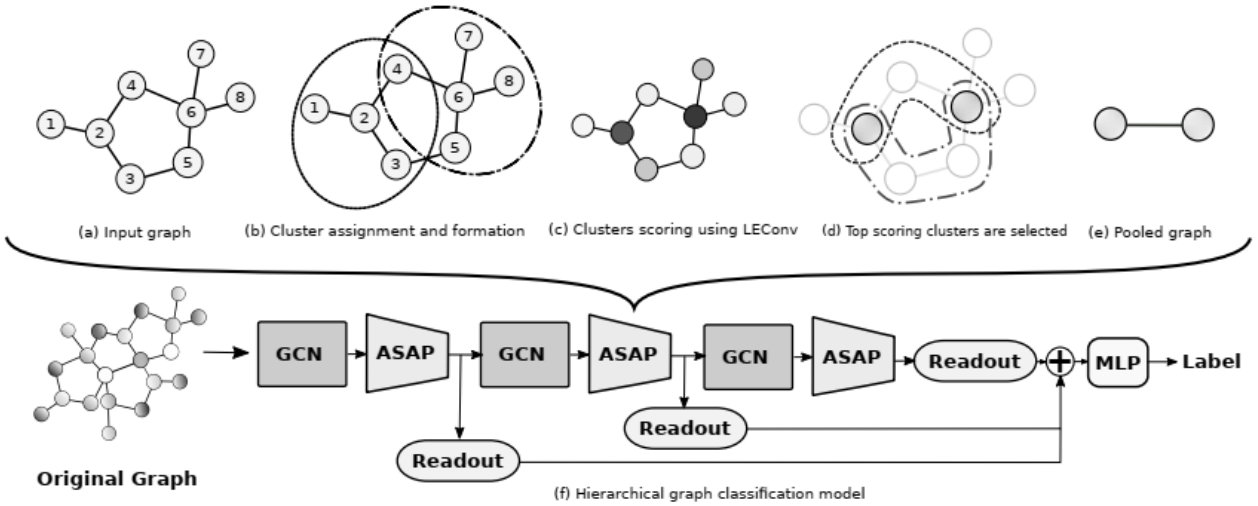
$$y(a_i, a_j) = \sigma(\beta n_{ij}) = \frac{1}{1 + \exp(-\beta n_{ij})}$$

- We search for the hyperparameter space and set $\beta = 0.5$ here

Trick 2: ASA Pooling

- DGCNN uses two elements by default:
  - Global Sort Pooling Layer
  - GCN Convolution Kernel

→ **Fast Overfitting**

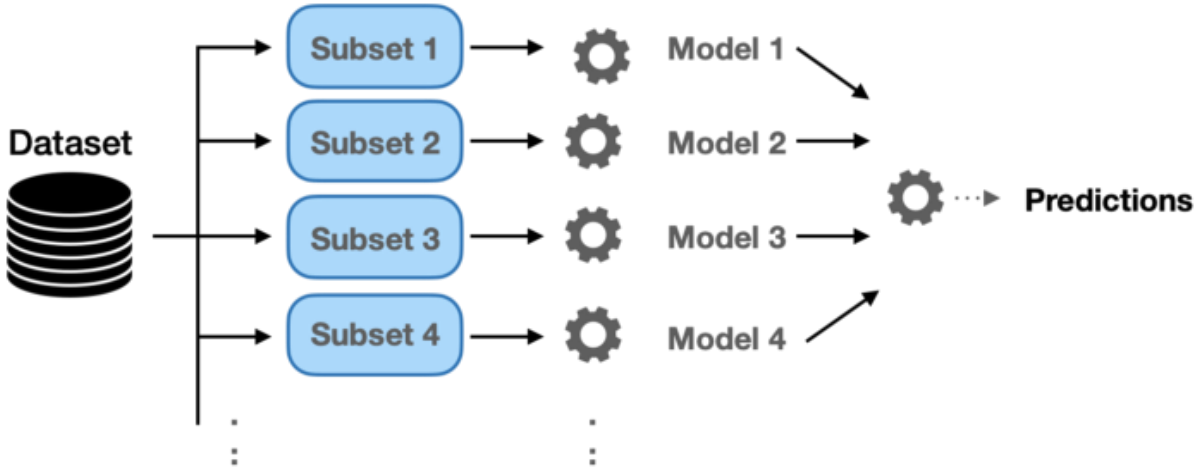


- Use Hierarchical ASAP structure, and substitute GCN conv. to LE conv.

E. Ranjan, S. Sanyal, et al. ASAP: Adaptive Structure Aware Pooling for Learning Hierarchical Graph Representations. AAAI 2020.

Enhance by Bagging

- Now we have obtain a single model which works very well
- However, we want to further improve our AUC score
- Common method in Kaggle competitions: **Ensemble Learning**



- About 0.01 ~ 0.02 AUC improvement

Summary and Acknowledgement

- We summarize our work as following four points:
  - We build a **unified homogeneous academic network**
  - We design a **simple but effective feature** for node classification
  - We improve the performance of SEAL by using **soft labels and ASAP**
  - We utilize **ensemble learning** to further raise the strength of model

- Acknowledgements:
  - ACK. to Prof. Jiaxin Ding and T.A. Bowen Zhang's insightful discussions
  - ACK. to other groups for great competition

Thanks!