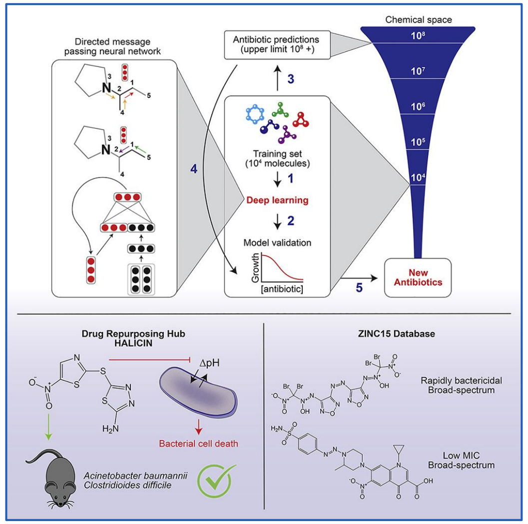
ETH zürich BIOMEDICAL INFORMATICS

Project 1
Training GNNs

Manuel Burger, Fedor Sergeev
Institute for Machine Learning,
Department of Computer Science

Paper presentations

- Starting next week
  - **Presenters**: share the slides on Monday
  - **Listeners**: skim the papers and attend
- Topic assignment
  - **40/50** selected a topic
  - All topics will be presented
  - Not all papers will be presented
- Paper assignment
  - In groups of 1 or 2 – it's fine
  - **10/40** did not select a paper
  - **No paper presentation – no credits !**
  - If you know your paper, but didn't register it on Moodle – **do it now**



[Stokes et al., A Deep Learning Approach to Antibiotic Discovery, 2020]

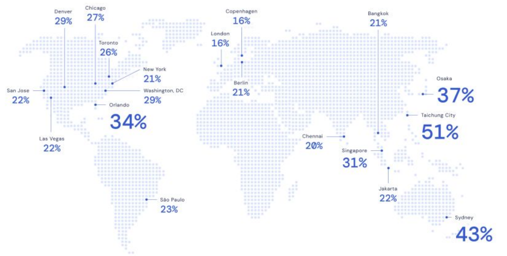


Figure 1: Google Maps estimated time-of-arrival (ETA) prediction improvements for several world regions, when using our deployed graph neural network-based estimator. Numbers represent relative reduction in negative ETA outcomes compared to the prior approach used in production. A negative ETA outcome occurs when the ETA error from the observed travel duration is over some threshold and acts as a measure of accuracy.

[Derrow-Pinion et al., ETA Prediction with Graph Neural Networks in Google Maps]

Projects

- **Format**
  - Team: 3 people
  - Assignment: Implement and apply deep learning on graphs
  - Coding: Python, PyTorch Geometric, Google Colab
  - Handouts (pdf) with point assignments will be provided
  - Deliverables
    - pdf report (background and results)
    - jupyter notebook (code and comments)
    - * presentation of notable projects
- **Project 1**
  - Duration: 3 weeks
  - Dates: out Oct 11 – due Nov 1
  - Topic
    - intro to GNNs and PyTorch geometric
    - node and graph level tasks
    - graph visualization and statistics

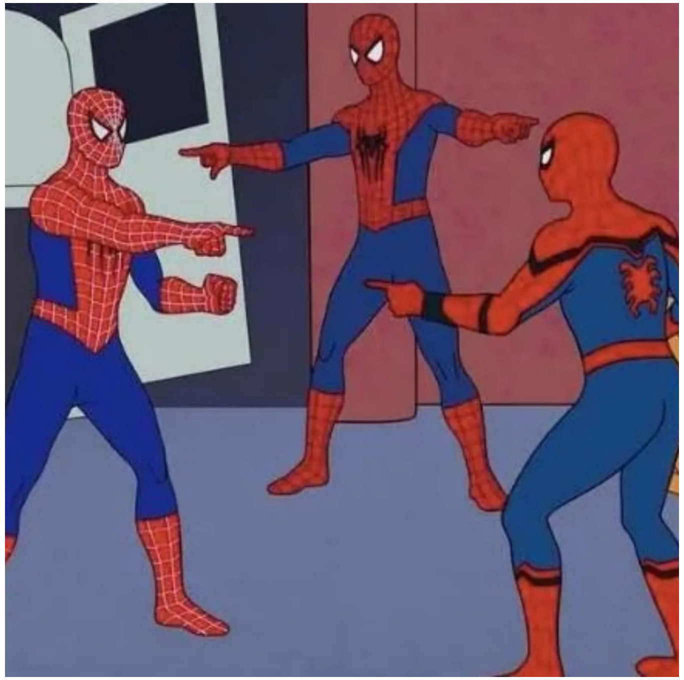
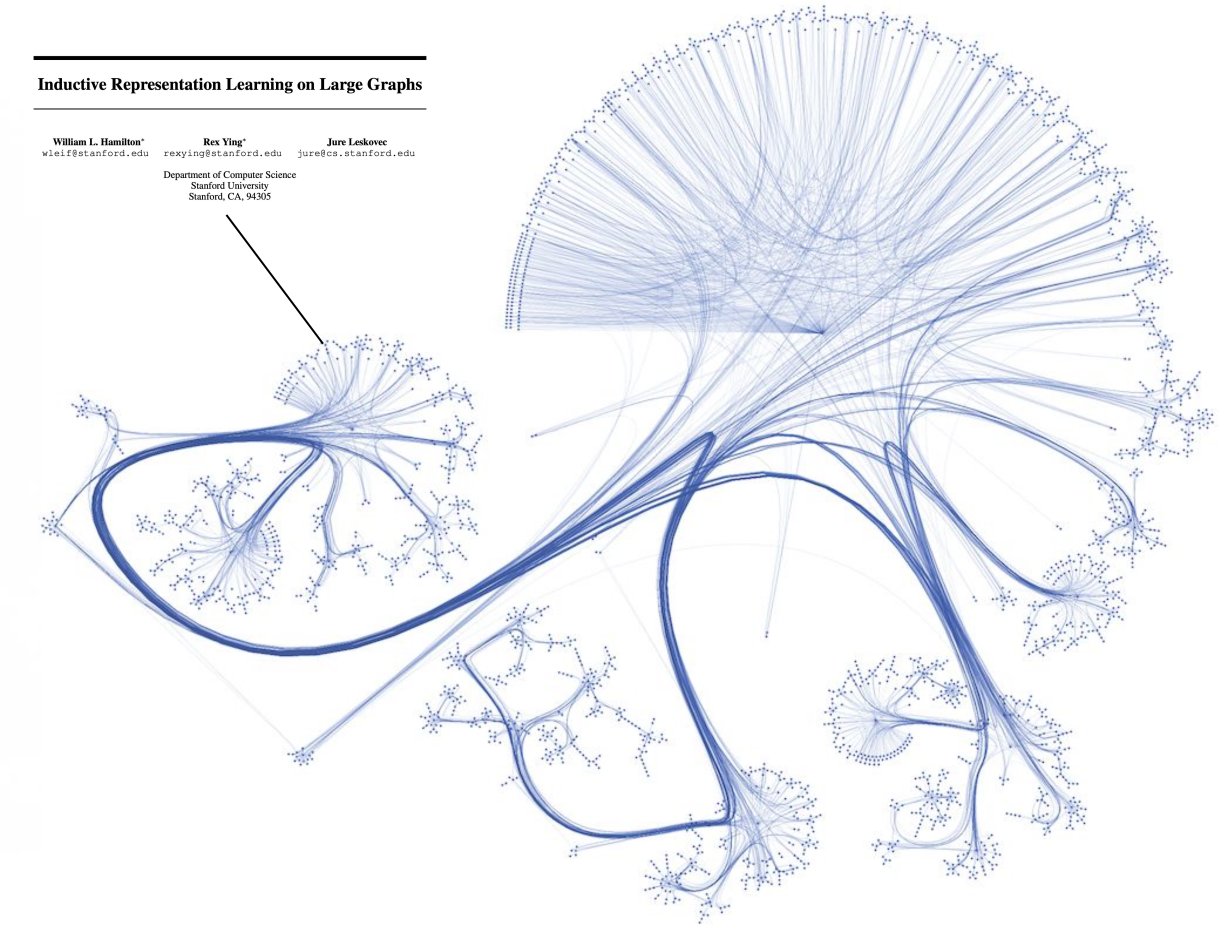


Fig.: Complete directed homogeneous graph with 3 vertices.

Part 2 - Cora Citation Dataset

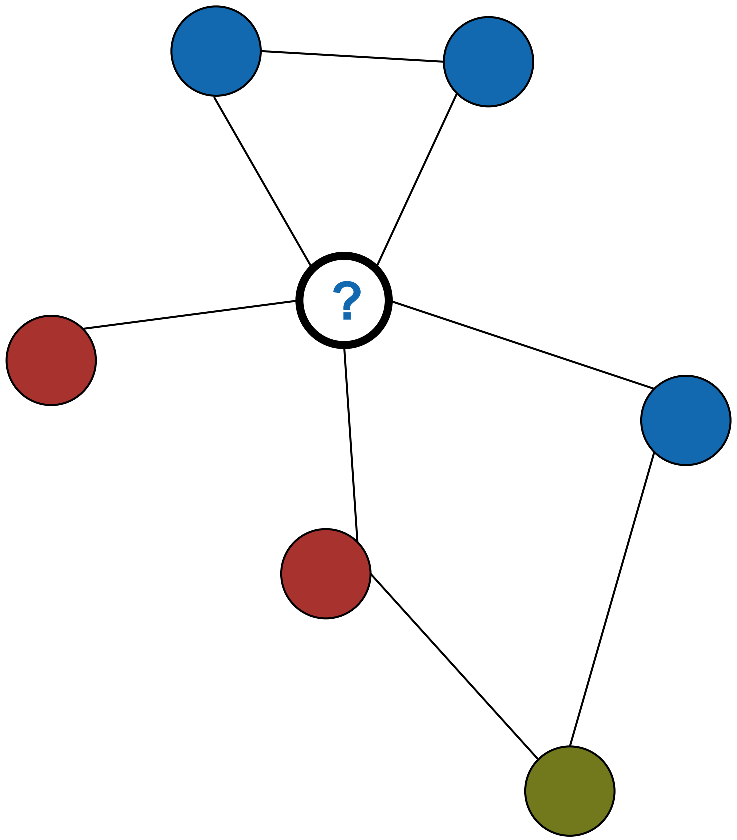
- Transductive Setting
- 2708 Scientific publications
- 5429 Citation Links
- 7 categories for classification
- Bag-of-Word feature vector
- `torch_geometric.datasets.Planetoid`
  - `Planetoid(name='Cora', split='public')`
- *Task 2.1: Data Exploration*



2.2 Label Propagation with full observations

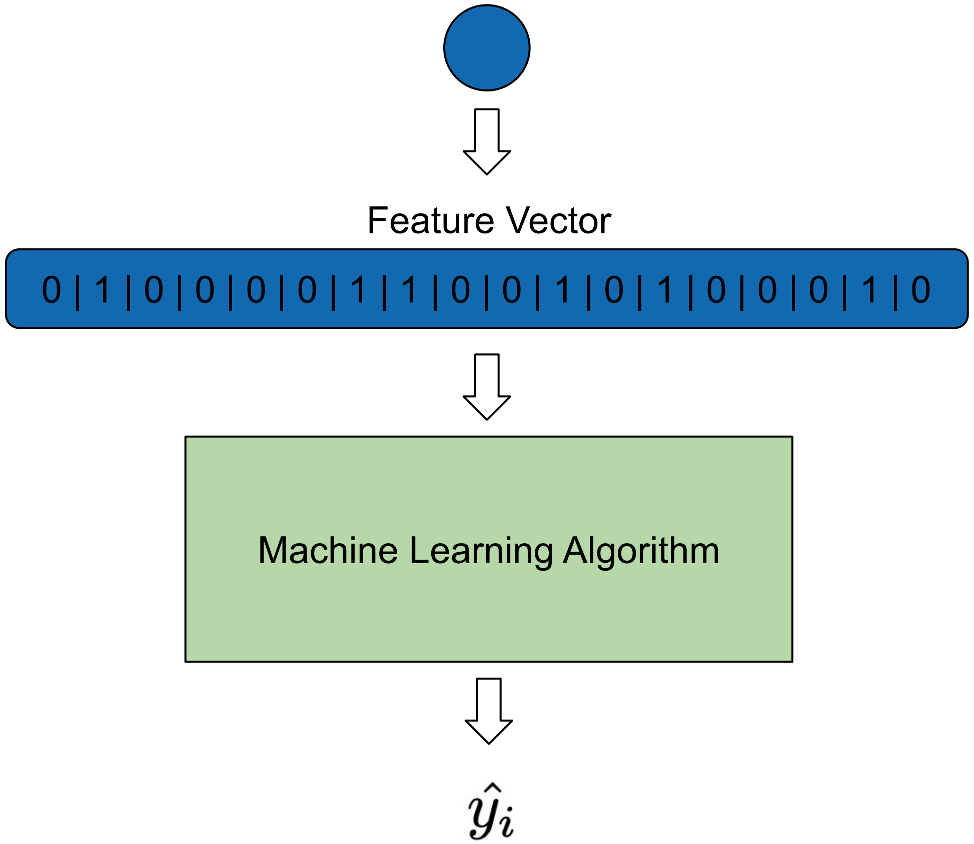
- What if we only learn from *structure*?
- For this task you are allowed to consider *all the labels, except for the current node* to be classified.
- We want to implement a *nearest neighbour predictor* solely based on the structure of the graph.

$$\hat{y}_i = Majority\left(\{y_j \mid j \in \mathcal{N}(i)\}\right)$$



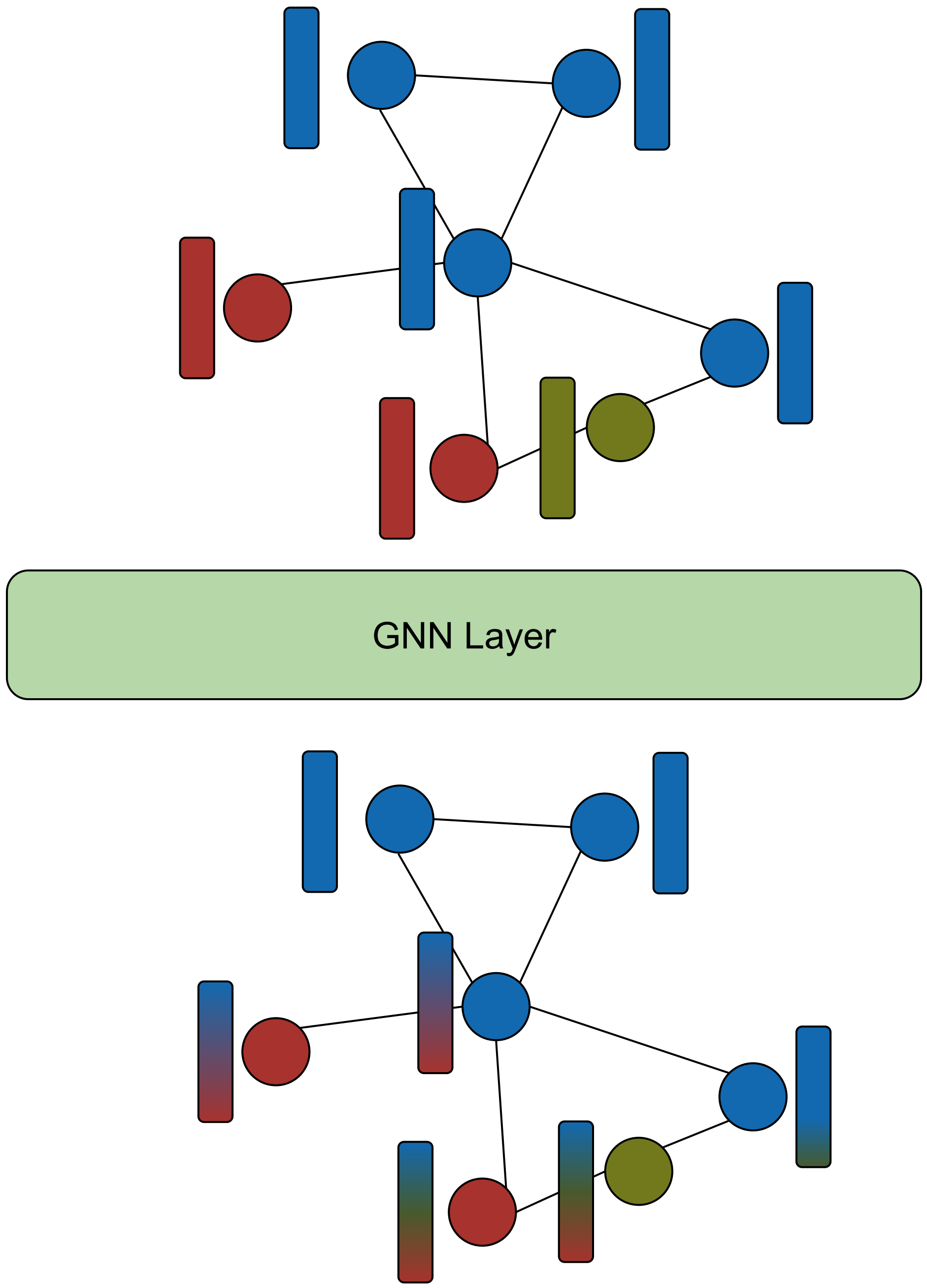
2.3 Baseline without Graph Structure

- What is the simplest model, which can solve our task? How much do we gain from sophisticated architectures?
- We establish a baseline by training a predictor on the node features



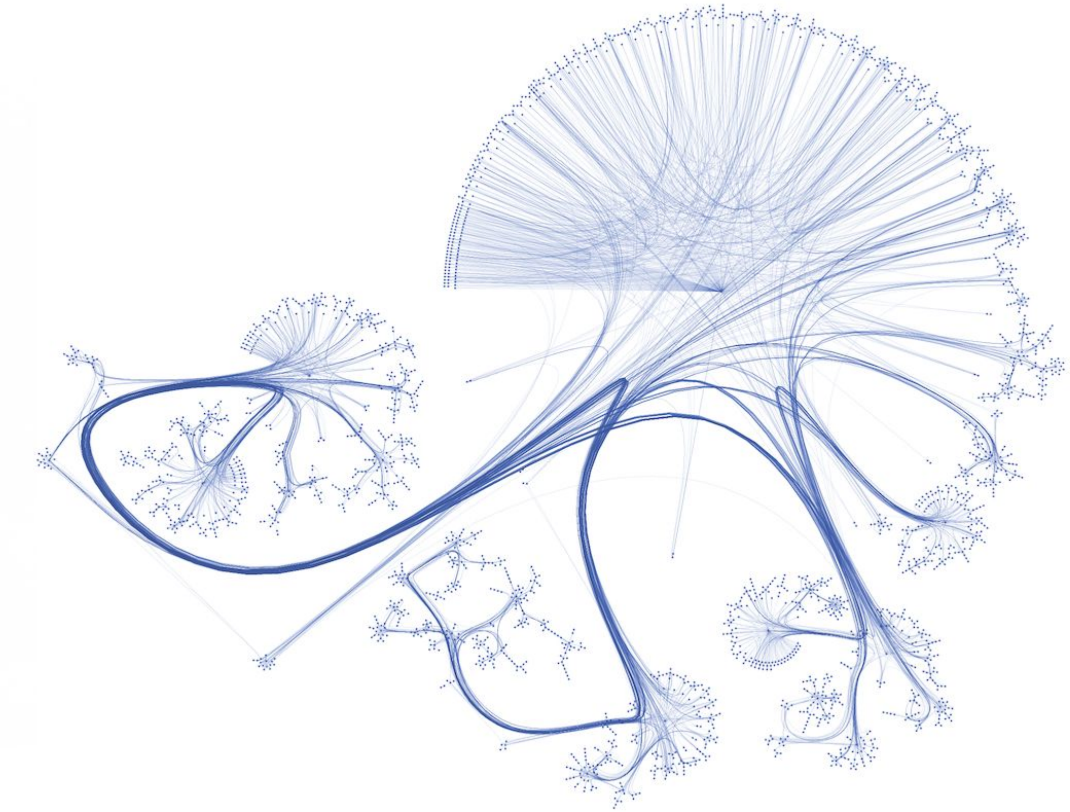
2.4 Untrained GNNs

- GNNs have a very powerful prior!
- Random projections can be used for dimensionality reductions
- Due to the nature of message passing GNNs with random initialization already exhibit interesting properties
- The goal of this task is to understand these properties:
  - Compare their performances
  - Visualizing embedding spaces



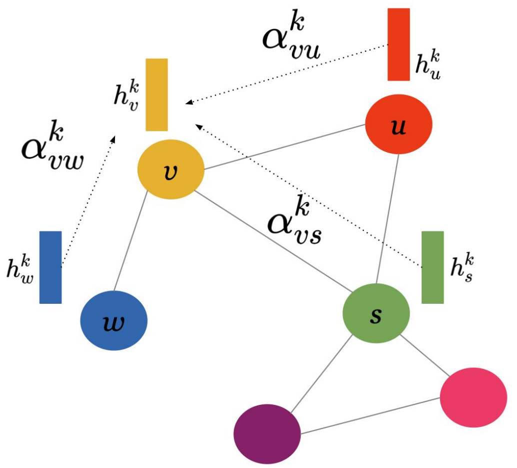
2.5 Trained GNNs

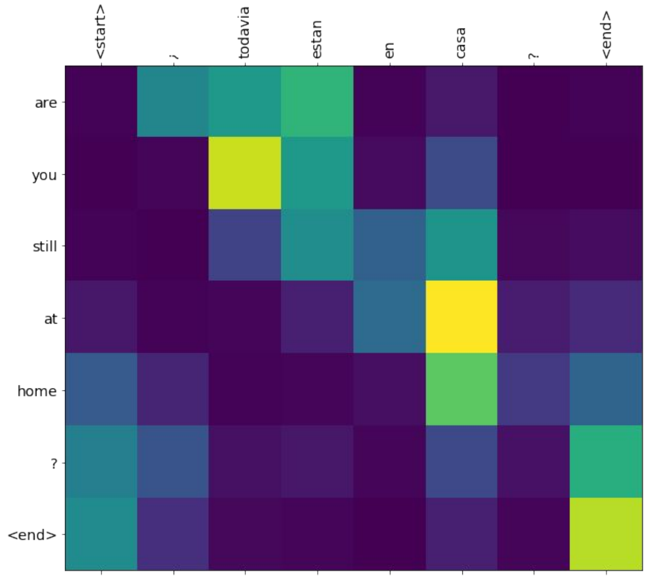
- Example Notebook for different architectures
- Train the GNN and put it into perspective to previously obtained numbers
- We encourage you to play around with different ideas you might have on how to combine available PyTorch / Pytorch Geometric Layers



2.6 Visualizing Graph Attention Networks

- Attention has become a popular mechanism to investigate neural network computations.
  - Although a bit controversial recently
- Still, we can get an intuition and see if it aligns with our priors
- The goal, train and visualize a graph attention network:
  - Are the attention weights channeling information in a meaningful way?



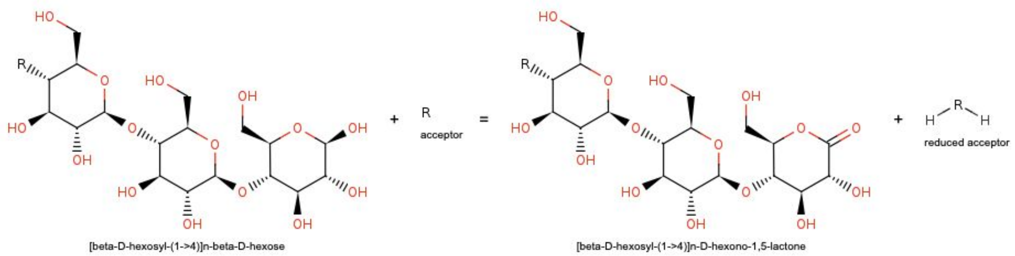


https://www.tensorflow.org/text/tutorials/nmt_with_attention

Part 3 - ENZYMES Dataset

- Inductive Setting: Many Graphs (600)
- Graph Classification
- 6 categories / classes (EC Number)
  - *“The Enzyme Commission number (EC number) is a numerical classification scheme for enzymes, based on the chemical reactions they catalyze” - Wikipedia*
- Small Notebook to explain advanced

  mini-batching on graphs



BRENDA

https://www.brenda-enzymes.org/index.php

Part 3 - ENZYMES Dataset

Implement a GNN with good classification accuracy:

- Baseline Performance: reach 40%
- Push Performance: reach 60% (get close)
- State of The Art: reach 70+% (Bonus)

Part 4 - Custom Message Passing

- Custom Implementation of GraphSAGE
- The notebook contains a skeleton to implement/subclass the `MessagePassing` class from `Pytorch Geometric`

$$h_v^{(l)} = W_l \cdot h_v^{(l-1)} + W_r \cdot AGG(\{h_u^{(l-1)}, \forall u \in N(v)\})$$

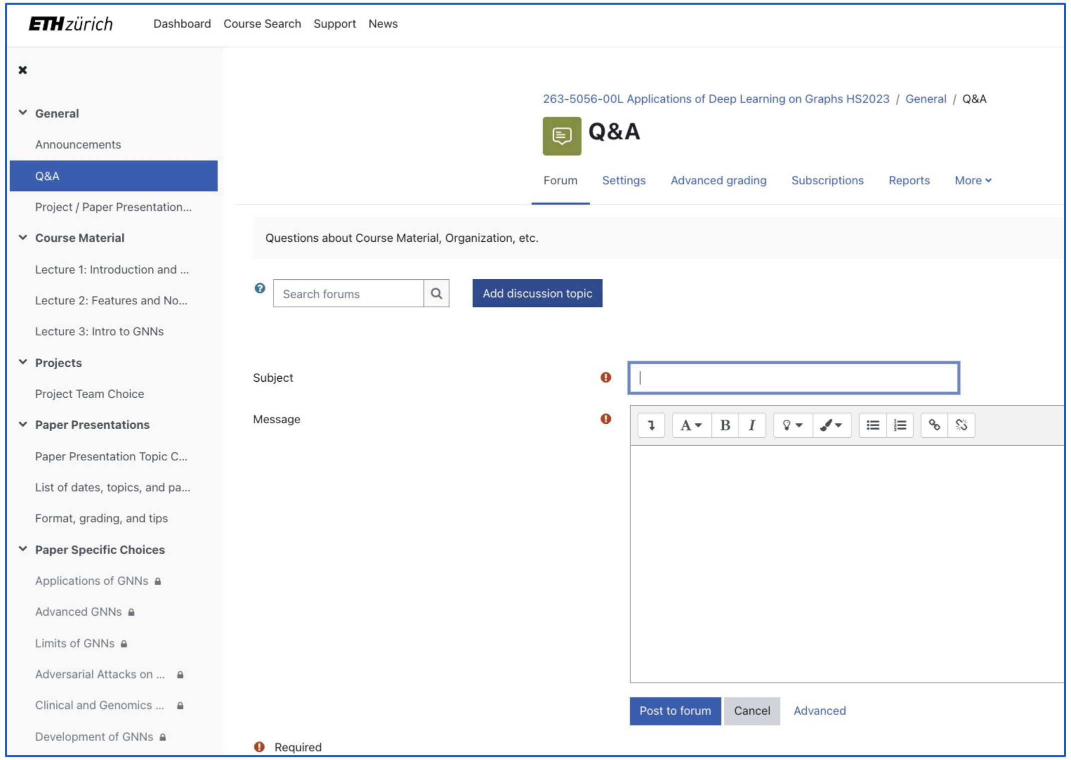
$$AGG(\{h_u^{(l-1)}, \forall u \in N(v)\}) = \frac{1}{|N(v)|} \sum_{u \in N(v)} h_u^{(l-1)}$$

Final thoughts

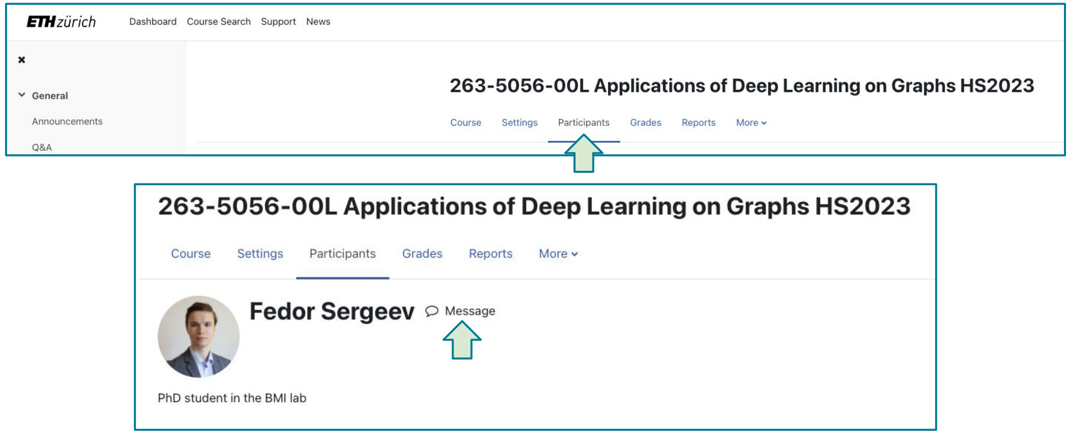
- Questions
  - Discuss within the group
  - Discuss between the groups
  - Ask on Moodle: in the Q&A forum or privately
- Please
  - Start early
  - Make the reports and notebooks easily readable
  - Check that the code runs
  - Don’t plagiarise (see Plagiarism Prevention)
- This is the 1st edition of the course, so there will be hiccups. Thank you for your patience!

Contact

- Through the Q&A forum
  a. Go to the ADL Moodle.
  b. Open the Q&A forum in the General section
  c. Click “Add discussion topic”
  d. Type in your question and “Post to forum”



- Privately
  a. Go to the ADL Moodle.
  b. Open participants
  c. Find a TA (Fedor Sergeev or Manuel Burger)
  d. Click on the name
  e. Click on “Message”



BIOMEDICAL
INFORMATICS