Master Project

Open Graph Benchmark - Large Scale Challenge @ KDD Cup 2021

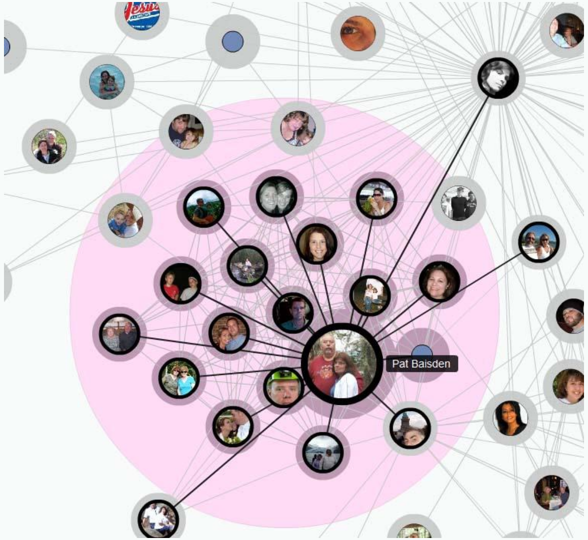
Supervised by: **Christian Beth**

Team members

1. Md. Mashiur Rahman
2. Ankit Malhotra
3. Abdullah Al Amin
4. Shaokat Hossain
5. Faiz Ahmed
6. Mithun Das
7. Md Abu Noman Majumder
8. Rishabh Lakra

Scope

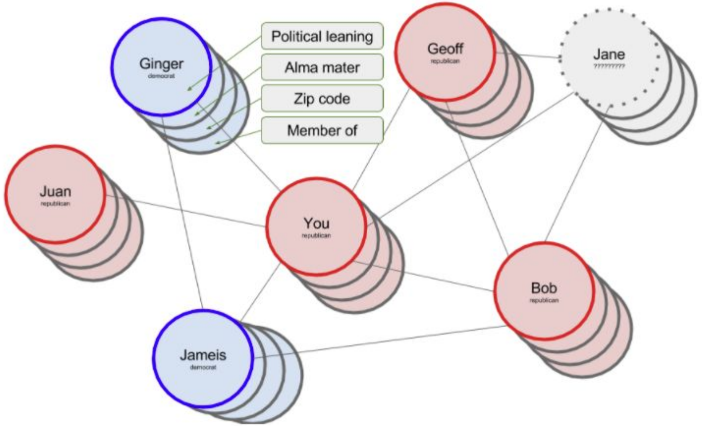
1. Graph data is everywhere
   a. Example: **Facebook** (Friend to Friend)
2. **GNN** is a powerful ML tool.
3. State-of-the-art models
4. **OGB** has a collection of large scale graphs.



Source:
http://www.messersmith.name/wordpress/wp-content/uploads/2009/10/social_graph_eunice_familly_cluster.jpg?fbclid=IwAR0ZZnUqQW1uMDWC1N2RADhjaoHGRAbPUUj0XIA6hFtBqbzvId8oHqxBpYo

Goal

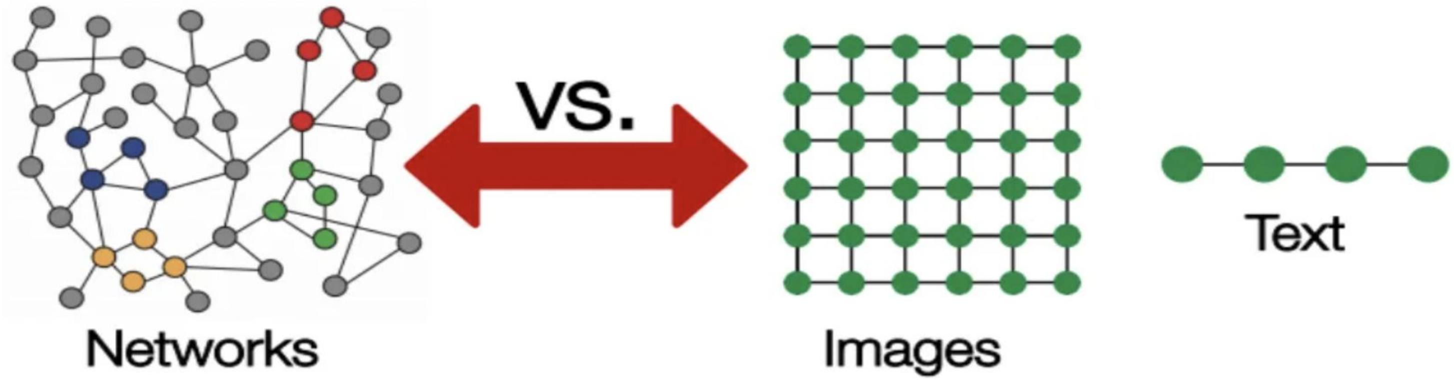
1. Learning about graph neural network
2. **KDD** Cup 2021
3. Apply **GNN** on heterogeneous graph



Source: https://www.experoinc.com/post/node-classification-by-graph-convolutional-network

From Image Convolution to Graph Convolution

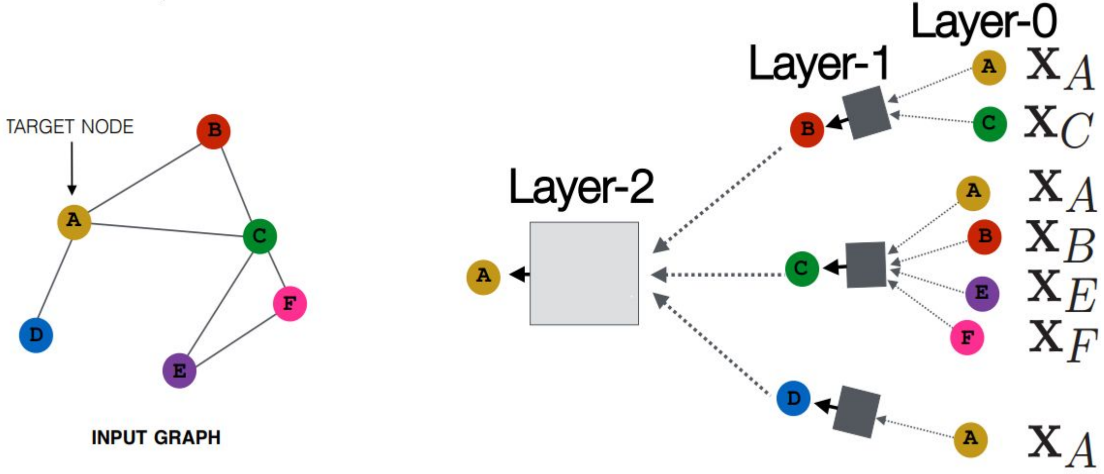
1. Image can be expressed as a **regular graph**
2. Difficult to perform CNN on graph
   a) **Arbitrary size** of graph
   b) **Complex** topology



Source: https://docs.google.com/presentation/d/1c3EYrhxlbx-umVaumCCohonxeagxY98akYmC-FNahNs/edit#slide=id.gb7f755f824_0_86

Graph Neural Networks

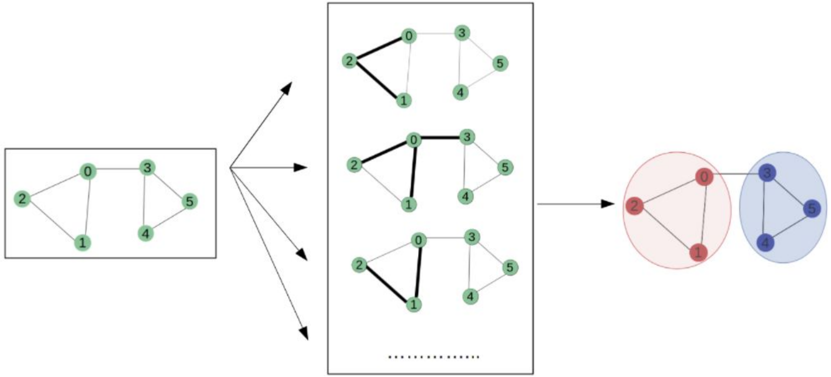
1. Each node contains embedded neighborhood information
2. Common tasks: Node labelling, node prediction, edge prediction, etc.
3. Convert edges by adding feed forward neural network layers and combine graphs and neural networks.



Source: https://docs.google.com/presentation/d/1c3EYrhxlbx-umVaumCCohonxeagxY98akYmC-FNahNs/edit#slide=id.gb7f755f824_0_92

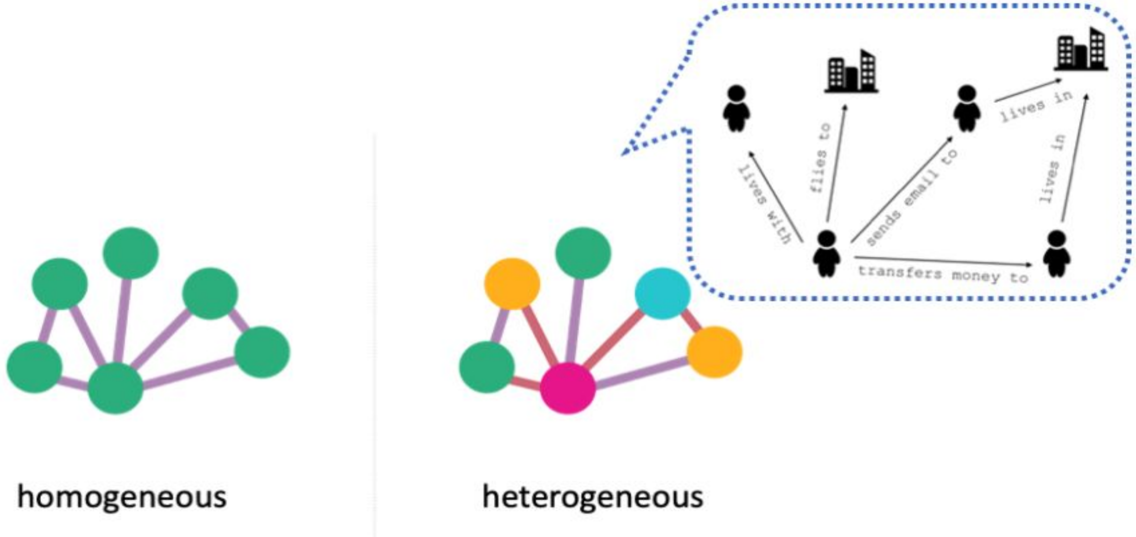
Graph Convolutional Networks (GCN)

1. A CNN that can work directly on graphs and leverage their structural information
2. ***Gather feature information → average → feed the average values into a neural network***
3. The number of layers is the farthest distance that node features can travel.



Source: https://docs.google.com/presentation/d/1c3EYrhxlbx-umVaumCCohonxeagxY98akYmC-FNahNs/edit#slide=id.gdc9da20b96_0_36

Homogeneous & Heterogeneous



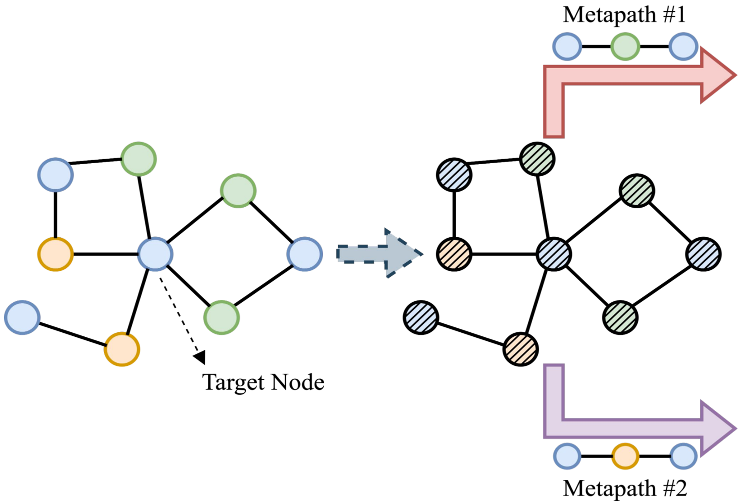
Source: https://medium.com/stellargraph/knowing-your-neighbours-machine-learning-on-graphs-9b7c3d0d5896

A single type of node and a single type of edge

Two or more types of nodes and/or two or more types of edges

Metapath

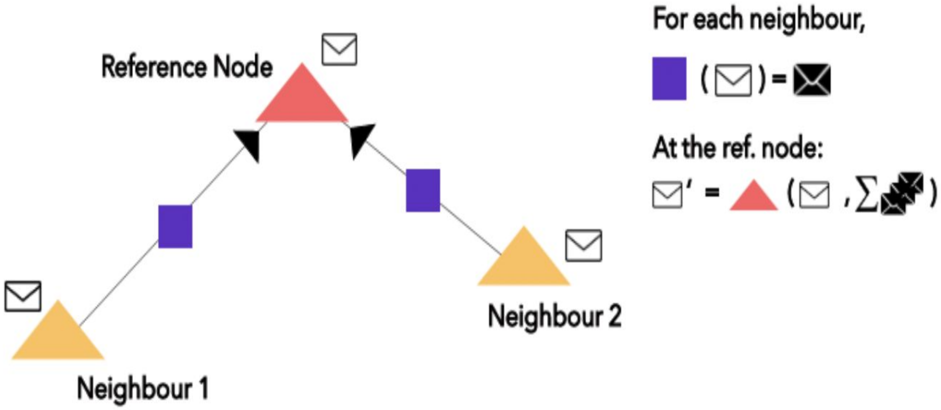
1. A composite relation connecting two objects
2. Widely used structure to capture the semantics



Source: https://deepai.org/publication/magnn-metapath-aggregated-graph-neural-network-for-heterogeneous-graph-embedding

Message passing framework

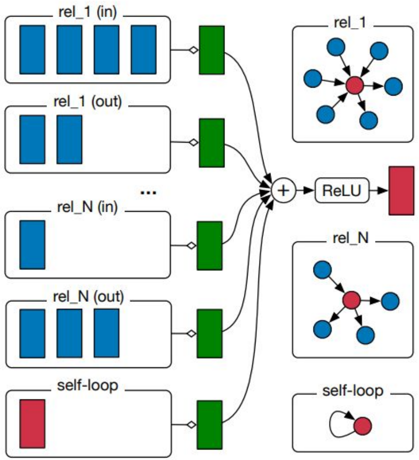
“These methods are based on some form of message passing on the graph, allowing different nodes to exchange information.”
-Michael Bronstein



Source: https://docs.google.com/presentation/d/1c3EYrhxlbx-umVaumCCohonxeagxY98akYmC-FNahNs/edit#slide=id.ge255dbf7a9_0_3

R-GCN

An effort to generalize GCN to handle different relationships between entities in a knowledge base.



GCN equation:

$$h_i^{l+1} = \sigma\left(\sum_{j \in N_i} \frac{1}{c_i} W^{(l)} h_j^{(l)}\right)$$

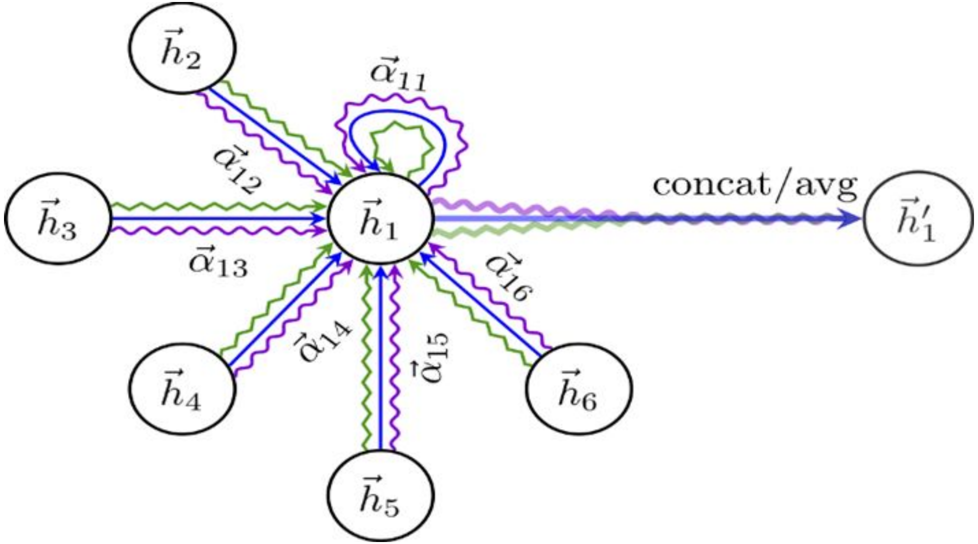
R-GCN equation:

$$h_i^{l+1} = \sigma\left(W_0^{(l)} h_i^{(l)} + \sum_{r \in R} \sum_{j \in N_i^r} \frac{1}{c_{i,r}} W_r^{(l)} h_j^{(l)}\right)$$

Source: https://docs.dgl.ai/en/0.4.x/tutorials/models/1_gnn/4_rgcn.html

Graph Attention(GAT) Network

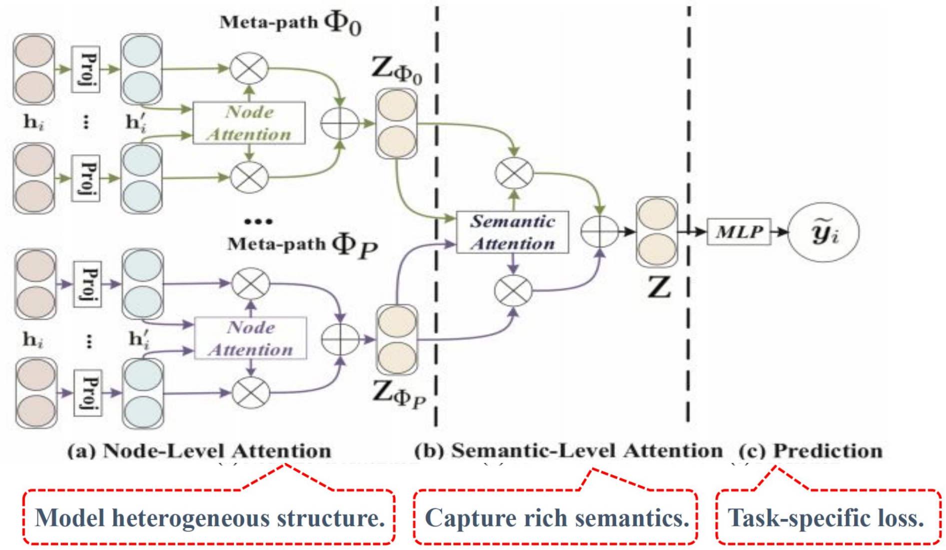
The GAT expands the basic aggregation function of the GCN layer.



Source: https://www.researchgate.net/figure/Schema-Metapath-and-Metagraph_fig1_323627097

HAN

The heterogeneity and rich semantic information bring great challenges for designing a graph neural network for heterogeneous graph.



Source: https://www.researchgate.net/figure/Schema-Metapath-and-Metagraph_fig1_323627097

Timeline

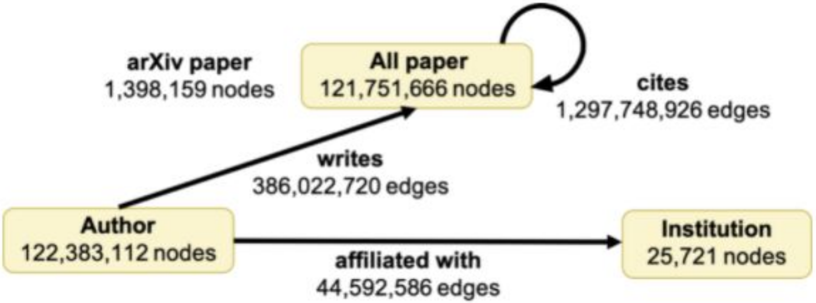
Week	Task
1-3	Paper reading
4th	Homo/Heterogeneous graph
5-6	MAG240M-LSC Dataset
7th	OGBN-MAG Dataset
8-9	ACM Dataset

Datasets

1. MAG240M-LSC (200 GB)
2. OGBN-MAG (1GB)
3. ACM Dataset(20MB)

MAG240M-LSC Dataset

It is a heterogeneous academic graph extracted from the Microsoft Academic Graph (MAG) .



Source: https://docs.google.com/presentation/d/1c3EYrhxlbx-umVaumCCohonxeagxY98akYmC-FNahNs/edit#slide=id.gb7f755f824_0_6

Task: Predict the primary subject areas of the given arXiv papers, which multi-class classification problem.

Tasks on **MAG240M-LSC**

1. Similar dataset of 2 GB (OGBN-MAG (Processed for PyG))
2. Sub-sampling of Actual Dataset
3. Working on the sub-sampled graph

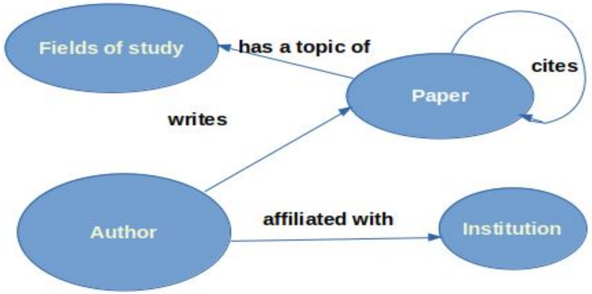
MAG240M-LSC Findings

1. Limitation of proper infrastructure
2. Sub-sampling a heterogeneous graph is not a trivial task
    a. Preservation of node degree distribution
    b. Representative of the actual graph

OGBN-MAG

1. It is a heterogeneous network composed of a subset of the Microsoft Academic Graph (MAG)
2. It contains four types of entities—papers , authors, institutions , and fields of study with 4 meta path "affiliated with", "writes" , "cites" and "has a topic of".

Task: Predict the venue (conference or journal) of each paper, given its content, references, authors, and authors' affiliations.



Tasks on **ogbn-mag**

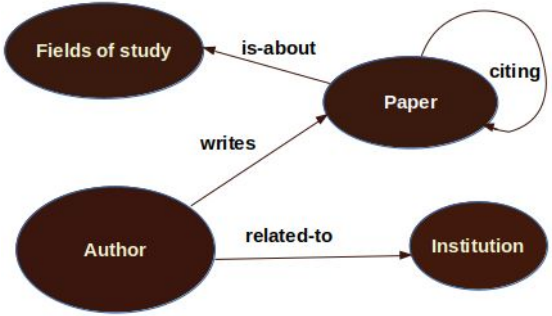
1. **RGCN**
   a. Prefilling missing feature with zero value
   b. Featureless Embedding
2. **HAN** Model with metapath

OGBN-MAG findings

1. **RGCN**(prefilled features) achieve 28% test accuracy
2. **RGCN**(featureless embedding) consumes full memory in paperspace after 20 epochs.
3. HAN **didn’t work** because of the DGL Library limitation

ACM Dataset

1. The **ACM dataset** is a heterogeneous network.
2. Entities - **papers** , **authors**, **institutions** and **fields of study**
3. Relation - **written-by**, **citing**, **is-about**, **published-in**, **related-to**, **contains** , **consist-of**.



Task: Predict the conference of a paper.

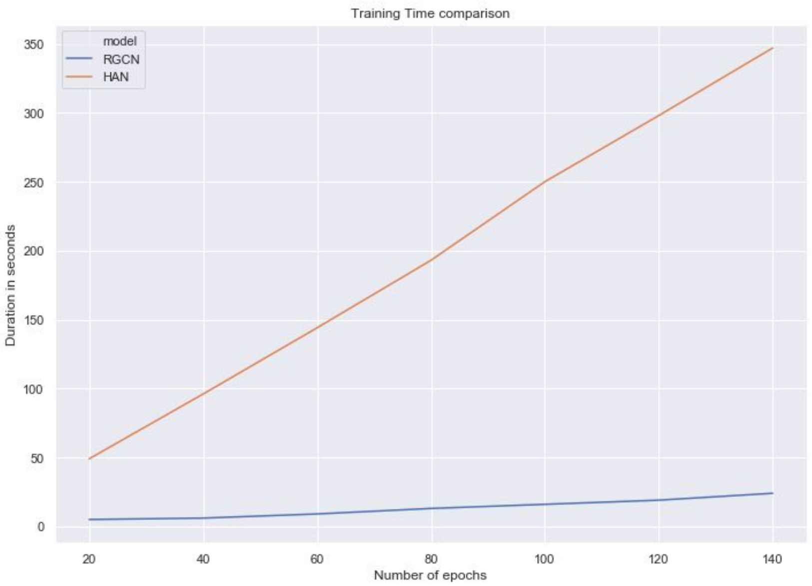
Tasks on **ACM Dataset**

1. **RGCN**
    a. Featureless embedding
2. **HAN** with 2 metapath
    a. PAP(paper-author-paper)
    b. PP(paper-paper)

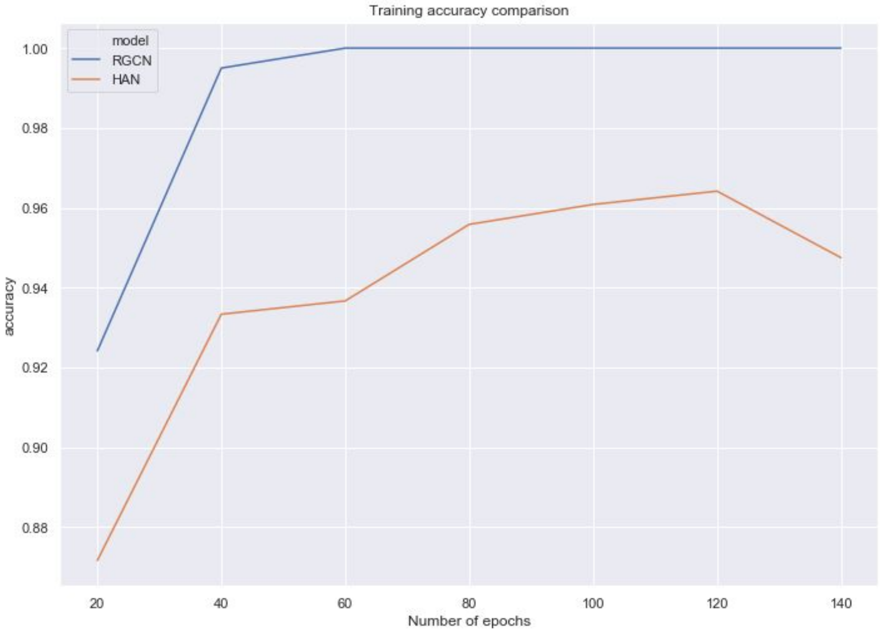
ACM findings

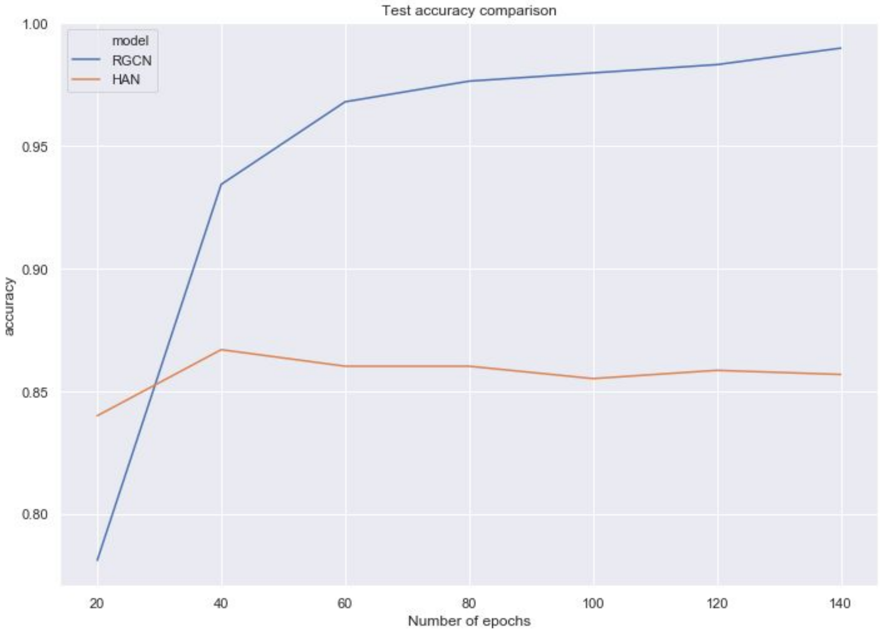
1. Capable of fitting both models.
2. **RGCN** shows better result than **HAN**
3. **HAN** takes much more time to train than **RGCN**

Training time:

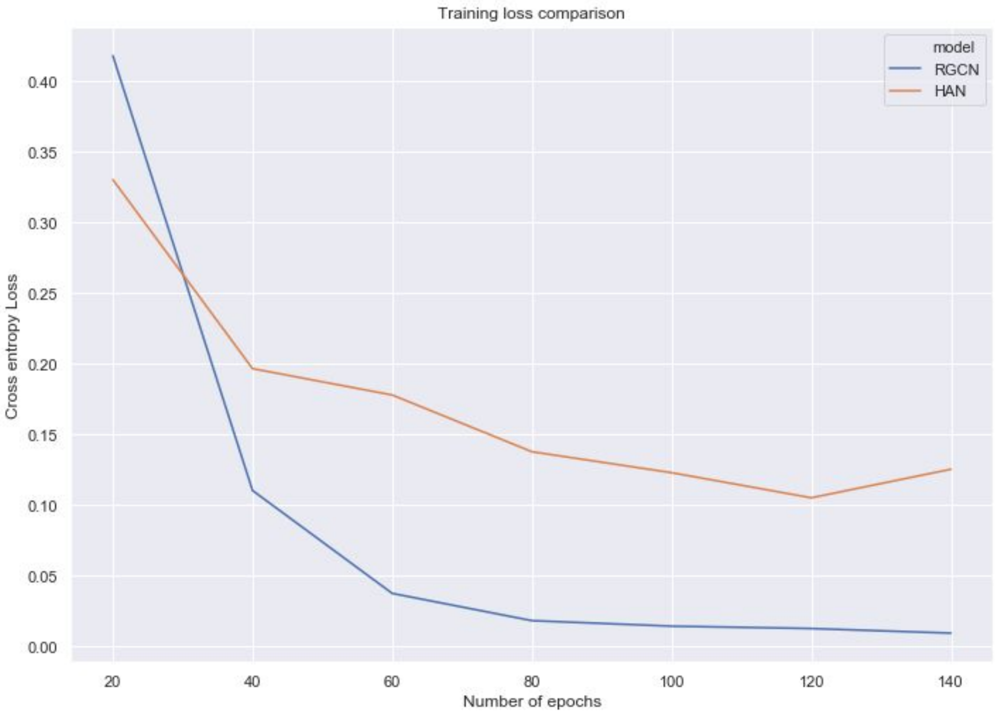


Accuracy:


Training accuracy comparison
model
RGCN
HAN
1.00
0.98
0.96
0.94
0.92
0.90
0.88
accuracy
20
40
60
80
100
120
140
Number of epochs


Test accuracy comparison
model
RGCN
HAN
1.00
0.95
0.90
0.85
0.80
accuracy
20
40
60
80
100
120
140
Number of epochs

Loss:


Training loss comparison
model
RGCN
HAN
Cross entropy Loss
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
20
40
60
80
100
120
140
Number of epochs

Challenges

1. Completely new concept
2. Resource limitation
3. Lack of proper documentation

Questions?

Thank you!