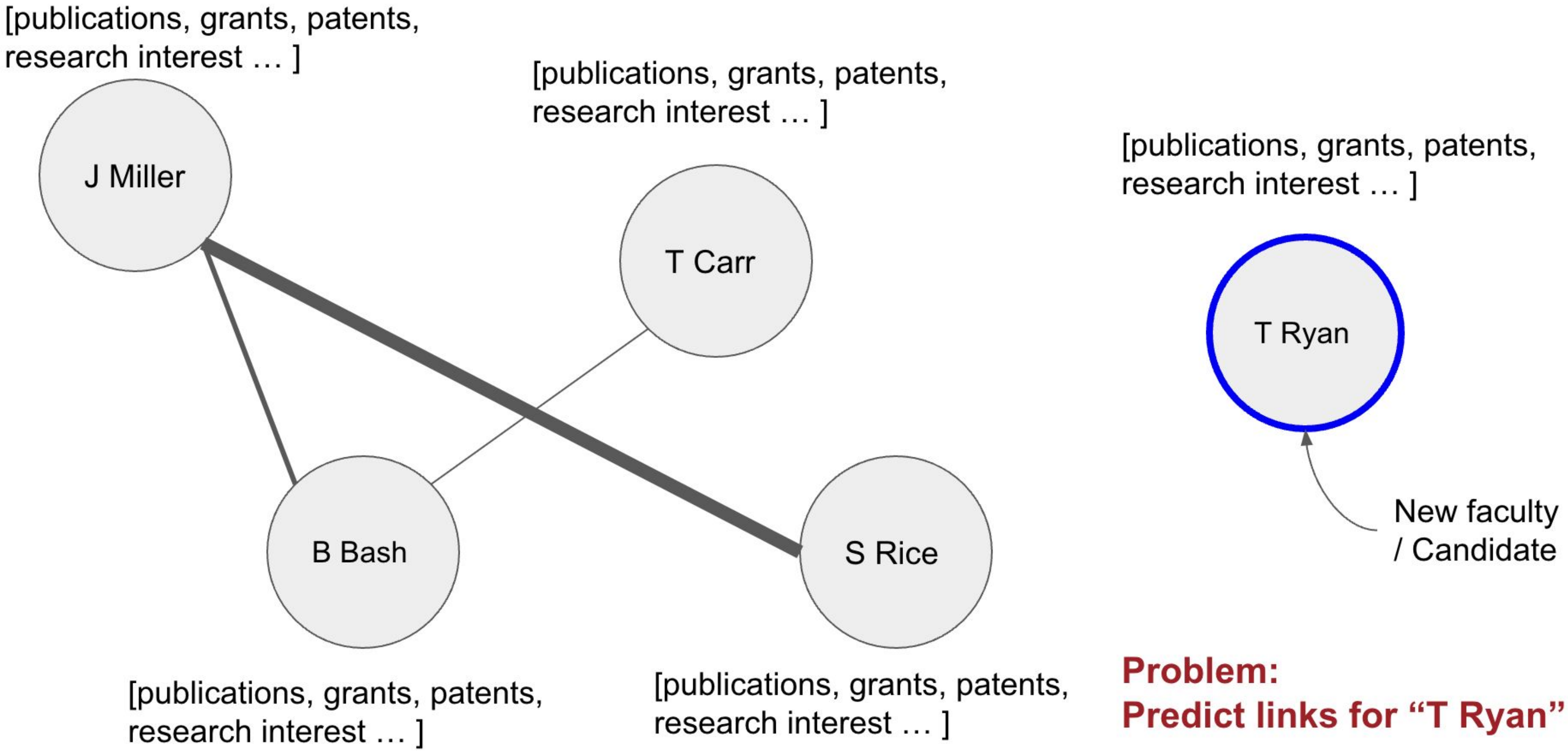
Link Prediction In Institutional Knowledge Graph

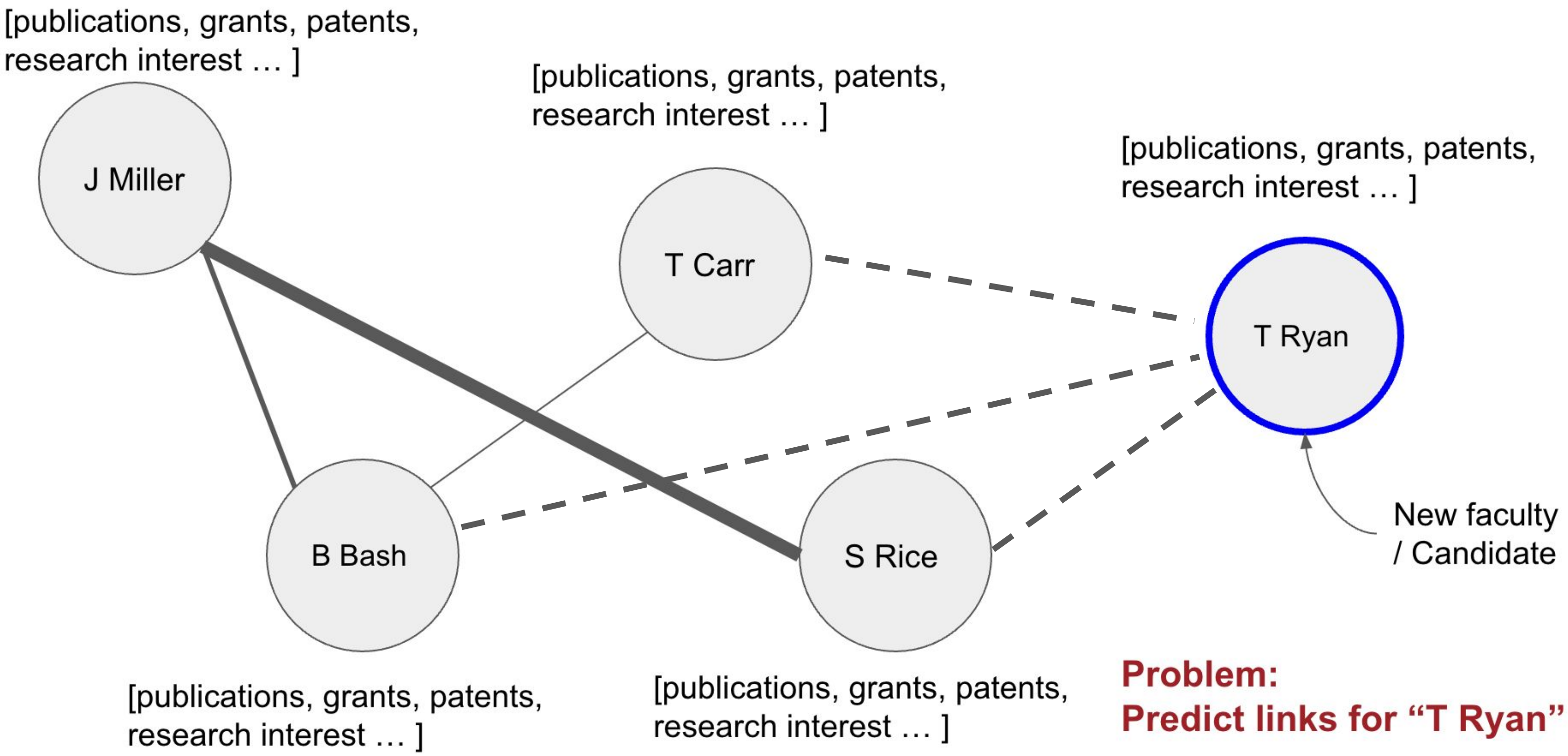
Sammi Abida Salma

Problem



Problem:
Predict links for "T Ryan"

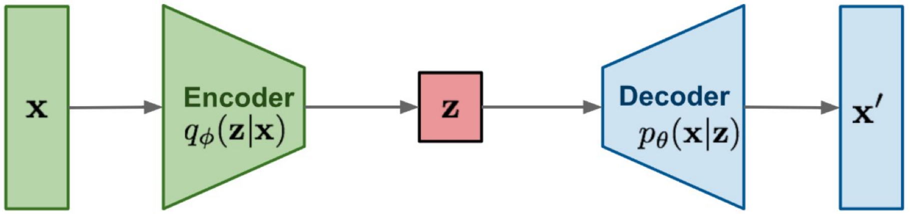
Link Prediction

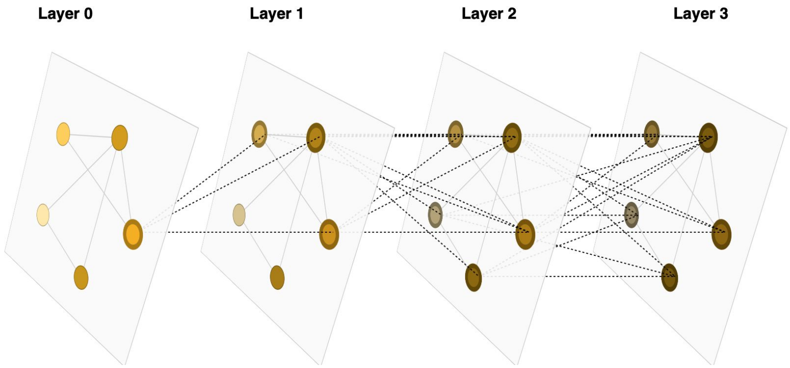


Problem:
Predict links for "T Ryan"

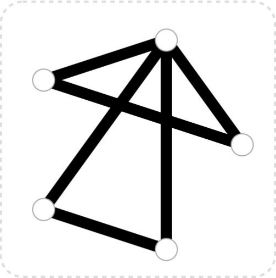
Approach

Link Prediction using Graph Auto-encoder





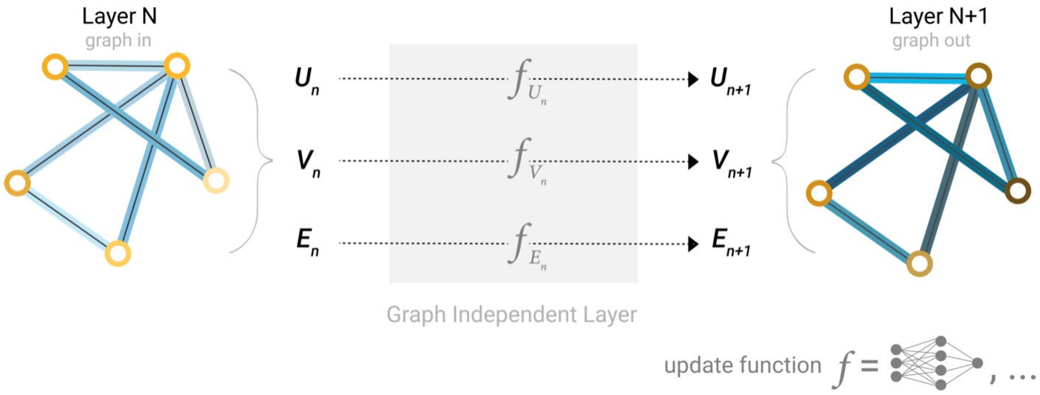
Graph Neural Network (GNN)



V Vertex (or node) attributes
e.g., node identity, number of neighbors

E Edge (or link) attributes and directions
e.g., edge identity, edge weight

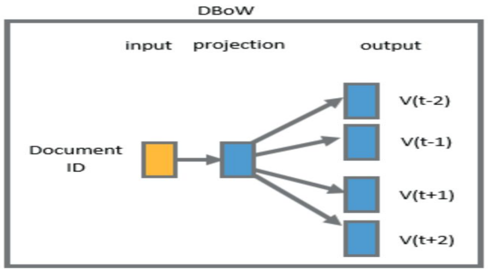
U Global (or master node) attributes
e.g., number of nodes, longest path



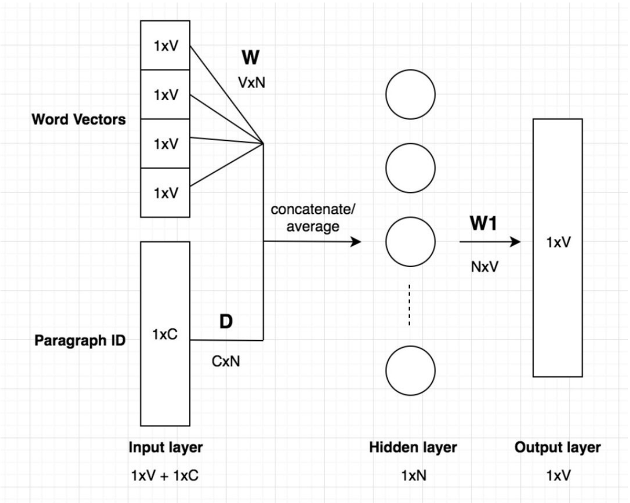
A single layer of a simple GNN. A graph is the input, and each component (V,E,U) gets updated by a MLP to produce a new graph. Each function subscript indicates a separate function for a different graph attribute at the n-th layer of a GNN model.

Doc2Vec

- **c**an predict the document's words based on its filename
- **k**nows which words go together in a document
- **u**ses the word similarities learned during training to construct a vector



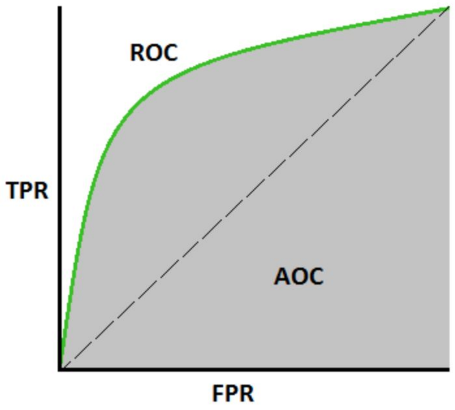
Distributed Bag-Of-Words Model



Evaluation

AUC - ROC Curve

Receiver Operator Characteristic (ROC)

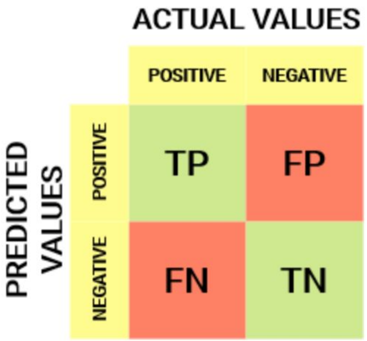


ROC plots the TPR against FPR at various threshold values

AUC measures the ability of a classifier to distinguish between classes

Higher is better

Confusion Matrix



$$\text{TPR / Recall / Sensitivity} = \frac{TP}{TP + FN}$$

$$FPR = \frac{FP}{TN + FP}$$

Case Study

Remove edges from the graph for case node

Estimate edges for that node

Compare with true edges

Experiment

Data Collection

Graph : collected in pickle format

Document (Titles of publications) : Collected publication titles from api

Preprocess data -> networkx graph data

Experiment Result

AUC ROC score = **0.9536788116320705**

Estimated # positive edges = 1040229

TP= 12399 FP= 1027830

FN= 60 TN= 3701217

TPR = 0.9951842041897424

FPR = 0.21734400186760672

Confusion Matrix

	True Positive	True negative
Estimated positive	TP= 12,399	FP= 1,027,830
Estimated negative	FN= 60	TN= 3,701,217

Actual # positive edges = 12,459
Actual # positive edges = 4,729,047

Experiment Result

# Layer	AUROC score	TP	FP
2	0.9536788116320705	12,399	1,027,830
3	0.9436318769052112	12,361	1,036,530
4	0.9559136142965436	12,371	1,020,674

Actual # positive edges = 12,459

Relation between # of layers and performance is undefined

Impact of Ratio = |Negative edges| : |Positive edges|

Ratio	Test performance (100 epoc)
1	0.9596
2	0.9576
3	0.9552
4	0.9564
5	0.9552
8	0.9524
10	0.9404
15	0.9394
20	0.9374

Actual # of positive edges = 12,459

# of negative edges = 2178*2177 - 12459

= 4,729,047

Ratio = 4729047/12459 ~ 380

Case Study

	actual	**Estimated** (With actual edges)	**Estimated** (Without actual edges)
kobourov	39	39	26
msurdeanu	37	37	15
janebambauer	26	26	5

Suggestions ?

to deal high false positive

General