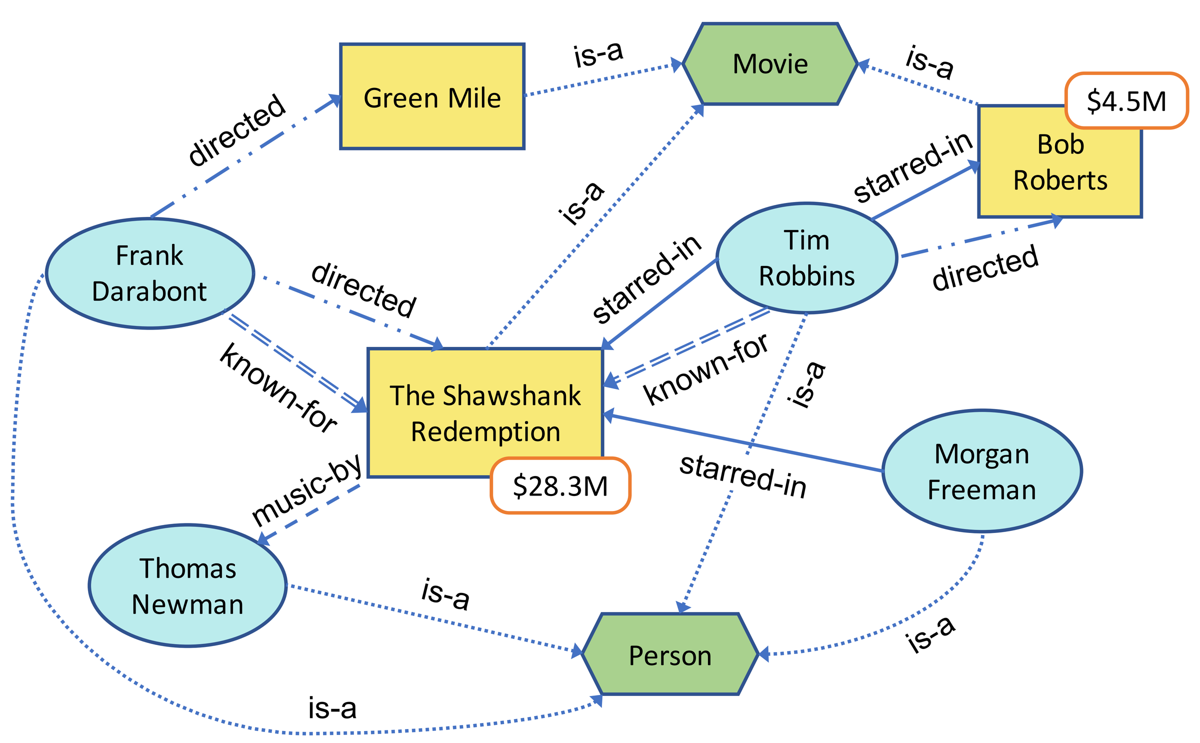
Estimating Node Importance in Knowledge Graphs Using Graph Neural Networks

KDD 2019

Namyong Park[1,2], Andrey Kan[2], Xin Luna Dong[2], Tong Zhao[2], Christos Faloutsos[1,2]
1. Carnegie Mellon University, 2. Amazon
Carnegie Mellon University Computer Science Department

Knowledge Graph

- A knowledge graph (KG) is a multi-relational graph representing facts in the form of “<subject> <predicate> <object>”
- Important for recommendation, Q/A, semantic search, etc
- Product Graph (Amazon), Freebase (acquired by Google), Satori (Microsoft), YAGO, DBpedia



How to Estimate Node Importance in KGs?

Problem Definition

Given a KG $G = (V, E = \{E_1, E_2, \ldots, E_p\})$ and importance scores $\{s\}$ for a subset $V_s \subseteq V$ of nodes, learn a function $S: V \to [0, \infty)$ that estimates the importance score of every in KG.

Desiderata for Modeling Node Importance in KGs

- *Neighborhood Awareness*
- *Making Use of Predicates*
- *Centrality Awareness*
- *Utilizing Input Importance Scores*
- Flexible Adaptation

Method Comparison

	GENI	HAR	PPR	PR
Neighborhood	✓	✓	✓	✓
Predicate	✓	✓		
Centrality	✓	✓	✓	✓
Input Score	✓	✓	✓	
Flexibility	✓			

Proposed Method: GENI

Overview

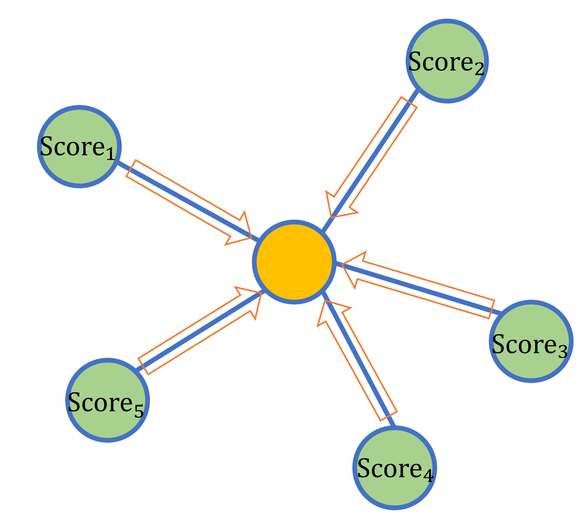
- GENI is a semi-supervised graph neural network (GNN)-based method that learns node importance
- GENI satisfies the above requirements

Score Aggregation

Models the relationship between the importance of neighboring nodes

$$s^{\ell}(i) = \sum_{j \in \mathcal{N}(i) \cup \{i\}} \alpha_{ij}^{\ell} s^{\ell-1}(j)$$

$$s^0(i) = \text{ScoringNetwork}(\vec{z}_i)$$



Predicate-Aware Attention Mechanism

Models how predicates affect the importance of neighboring entities using self-attention mechanism

$$\alpha_{ij}^{\ell} = \frac{\exp\left(\sigma_a\left(\sum_m \vec{a}_{\ell}^T [s^{\ell-1}(i) \| \phi(p_{ij}^m) \| s^{\ell-1}(j)]\right)\right)}{\sum_{k \in \mathcal{N}(i) \cup \{i\}} \exp\left(\sigma_a\left(\sum_m \vec{a}_{\ell}^T [s^{\ell-1}(i) \| \phi(p_{ik}^m) \| s^{\ell-1}(k)]\right)\right)}$$

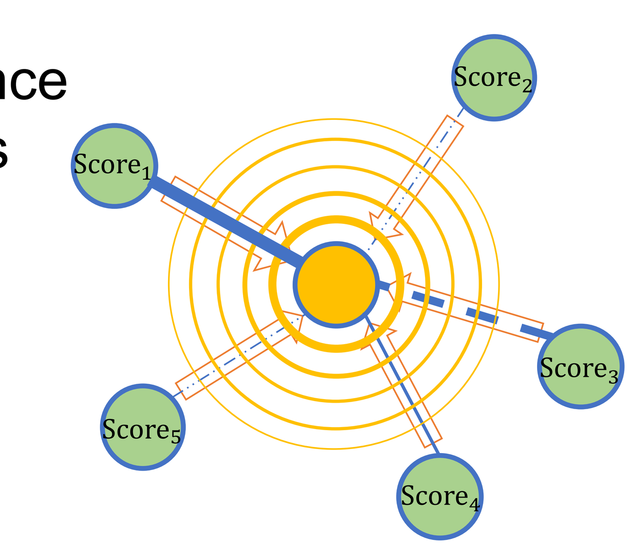
Centrality Adjustment

Makes use of the fact that the importance of a node normally positively correlates with its centrality in a network

$$c(i) = \log(d(i) + \epsilon)$$

$$c^*(i) = \gamma \cdot c(i) + \beta$$

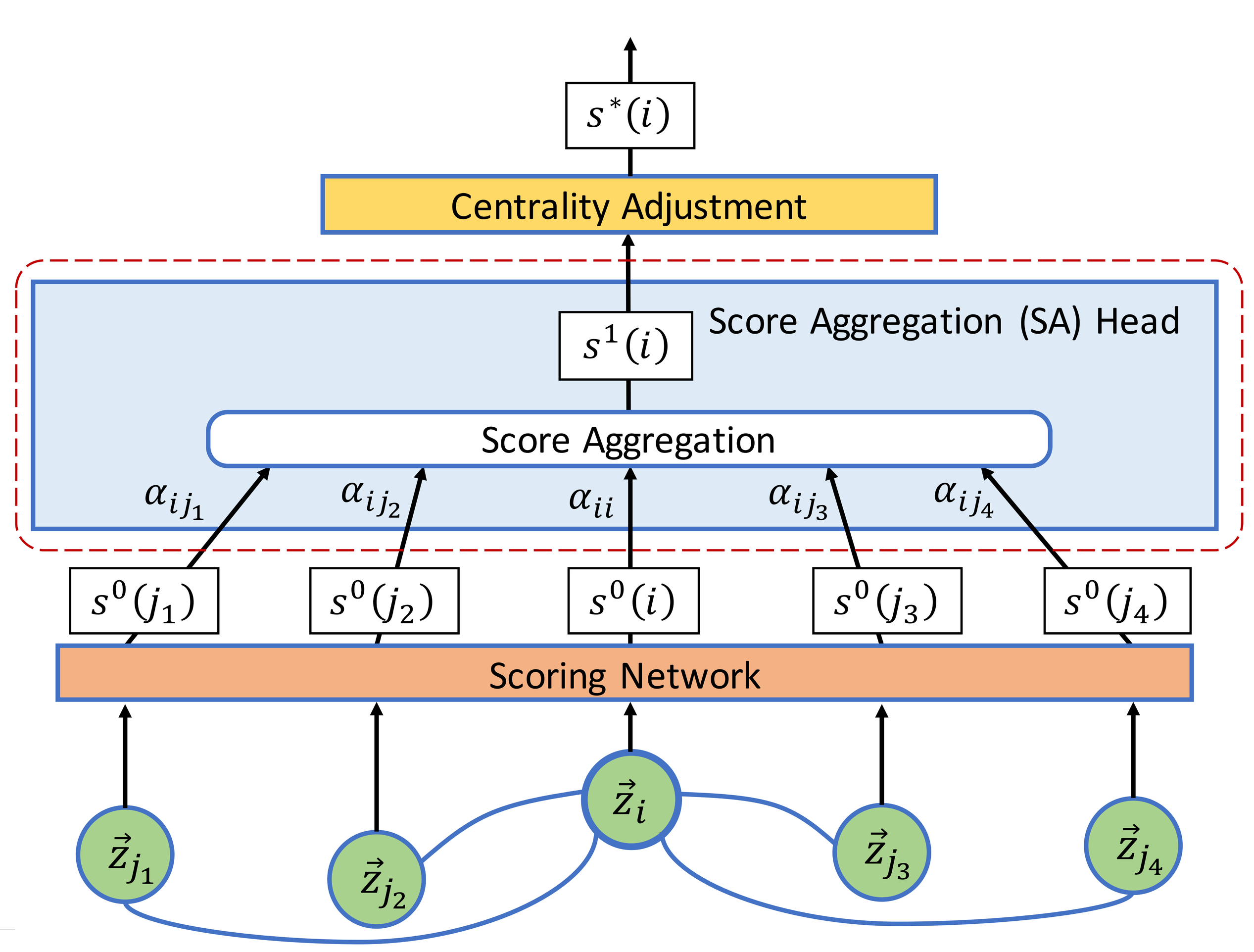
$$s^*(i) = \sigma_s(c^*(i) \cdot s^L(i))$$

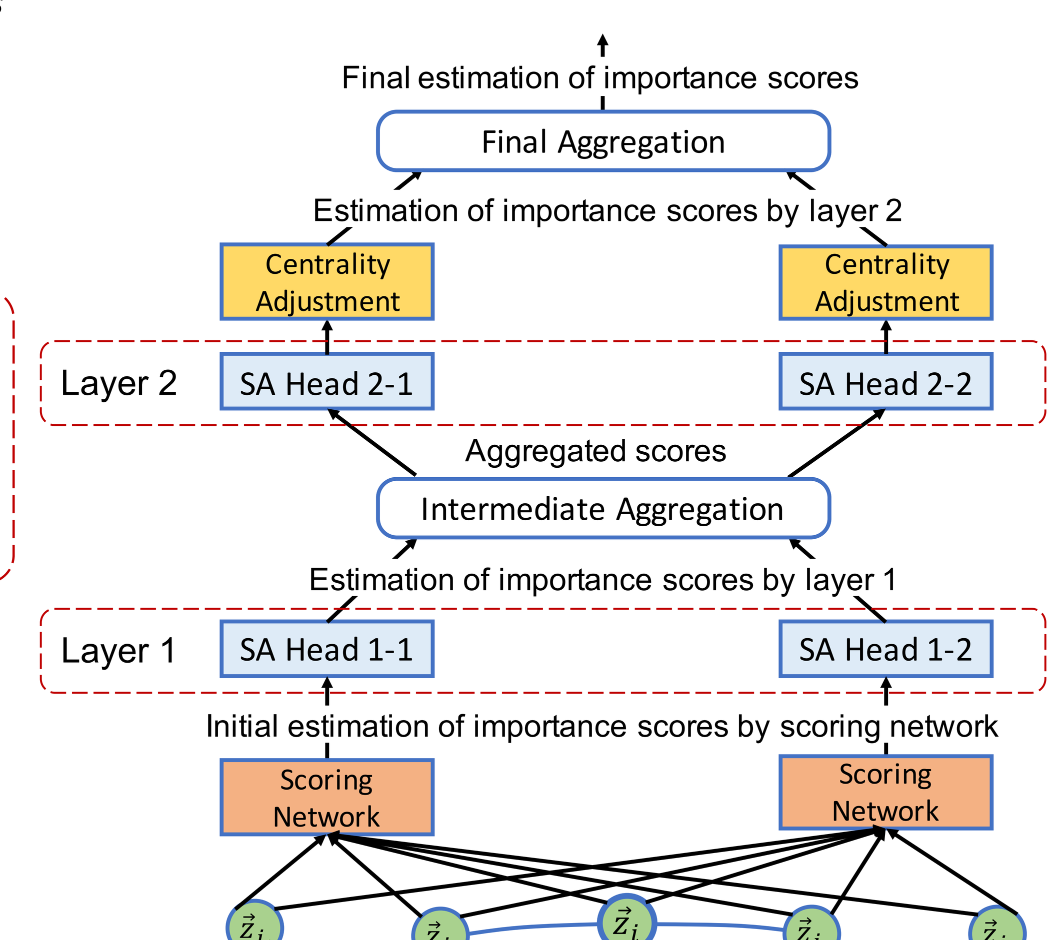


Model Training

$$L(\Omega) = \frac{1}{|V_s|} \sum_{i \in V_s} (s^*(i) - g(i))^2 + \lambda \|\Omega\|_2^2$$

GENI Architecture





Experiments

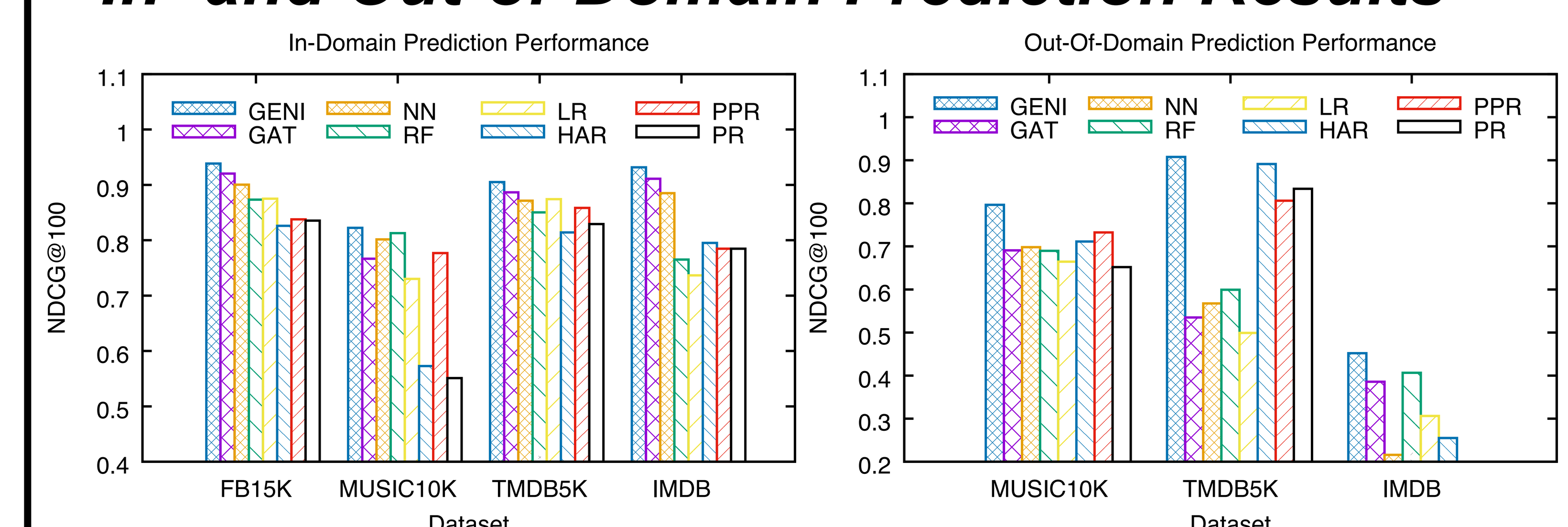
Datasets

Name	# Nodes	# Edges	# Predicates	Input Score Type	# Nodes w/ Scores	Data for OOD Evaluation
FB15K	14,951	592,213	1,345	# Pageviews	14,108 (94%)	N/A
MUSIC10K	24,830	71,846	10	Song hotttnesss	4,214 (17%)	Artist hotttnesss
TMDB5K	123,906	532,058	22	Movie popularity	4,803 (4%)	Director ranking
IMDB	1,567,045	14,067,776	28	# Votes for movies	215,769 (14%)	Director ranking

In-and Out-Of-Domain Evaluation

Given importance scores for some nodes $V_s \subseteq V$ of type $\mathcal{T}$ (e.g., movies), predicting the importance of nodes of type $\mathcal{T}$ is called an *“in-domain”* estimation, and importance estimation for those nodes whose type is not $\mathcal{T}$ is called an *“out-of-domain”* estimation.

In- and Out-of-Domain Prediction Results



*NN (Neural Network), LR (Linear Regression), PPR (Personalized PageRank), GAT (Graph Attention Networks), RF (Random Forests), HAR (Hub, Authority, and Relevance), PR (PageRank)

In-Domain Regression Performance

Method	FB15K	MUSIC10K	TMDB5K	IMDB
LR	1.3536 ± 0.017	0.1599 ± 0.002	0.8431 ± 0.028	1.7534 ± 0.005
RF	1.2999 ± 0.024	0.1494 ± 0.002	0.9223 ± 0.015	1.8181 ± 0.011
NN	1.2463 ± 0.015	0.1622 ± 0.009	0.8496 ± 0.012	2.0279 ± 0.033
GAT	1.0798 ± 0.031	0.1635 ± 0.007	0.8020 ± 0.010	1.2972 ± 0.018
GENI	**0.9471 ± 0.017**	**0.1491 ± 0.002**	**0.7150 ± 0.003**	**1.2079 ± 0.011**

Case Study on TMDB5K

Top-10 movies (in-domain estimation)

	GENI		HAR		GAT	
1	The Dark Knight Rises	11	Jason Bourne	63	The Dark Knight Rises	11
2	The Lego Movie	70	The Wolf of Wall Street	21	Clash of the Titans	103
3	Spectre	10	Rock of Ages	278	Ant-Man	4
4	Les Misérables	94	Les Misérables	94	The Lego Movie	68
5	The Amazing Spider-Man	22	The Dark Knight Rises	7	Jack the Giant Slayer	126
6	Toy Story 2	39	V for Vendetta	27	Spectre	7
7	V for Vendetta	26	Now You See Me 2	81	The Wolf of Wall Street	16
8	Clash of the Titans	97	Spectre	5	The 5th Wave	67
9	Ant-Man	-2	Austin Powers in Goldmember	140	The Hunger Games: Mockingjay - Part 2	-4
10	Iron Man 2	29	Alexander	141	X-Men: First Class	767

Top-10 directors (out-of-domain estimation)

	GENI		HAR		GAT	
1	Steven Spielberg	0	Steven Spielberg	0	Noam Murro	N/A
2	Tim Burton	9	Martin Scorsese	44	J Blakeson	N/A
3	Ridley Scott	6	Ridley Scott	6	Pitof	N/A
4	Martin Scorsese	42	Clint Eastwood	19	Paul Tibbitt	N/A
5	Francis Ford Coppola	158	Woody Allen	112	Rupert Sanders	N/A
6	Peter Jackson	-4	Robert Zemeckis	1	Alan Taylor	145
7	Robert Rodriguez	127	Tim Burton	4	Peter Landesman	N/A
8	Gore Verbinski	8	David Fincher	40	Hideo Nakata	N/A
9	Joel Schumacher	63	Oliver Stone	105	Drew Goddard	N/A
10	Robert Zemeckis	-3	Ron Howard	-2	Tim Miller	N/A

Conclusions

- We proposed GENI, a novel graph neural network that estimates node importance in knowledge graphs
- GENI shows superior performance on both in- and out-of-domain prediction tasks