Hyperlink Prediction on Hypergraphs of Text

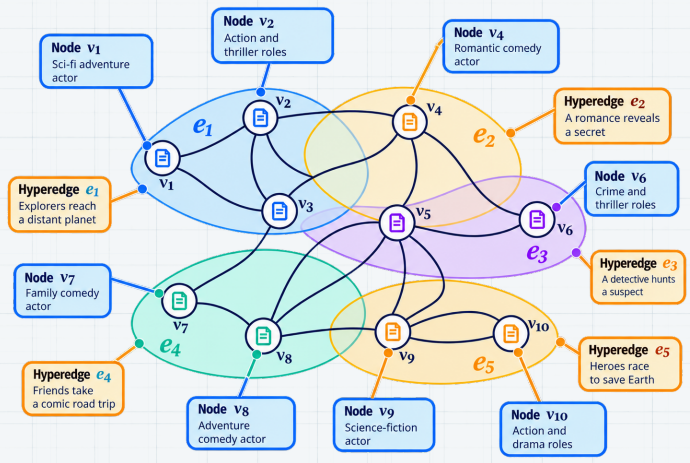
Alessia Antelmi[2] · Tiziano Citro[1] · Dario De Maio[1] · Daniele De Vinco[1], Valerio Di Pasquale[1] · Mirko Polato[2] · Carmine Spagnuolo[1]

UNIVERSITY OF SALERNO[1] / UNIVERSITY OF TURIN[2]

01 / THE MOTIVATION

Hypergraphs model many-to-many relations, but most hyperlink prediction methods focus mainly on the topology. In many applications, nodes and hyperedges also have text that provides context.

HYPERGRAPH OF TEXT



COMBINING STRUCTURAL AND SEMANTIC INFORMATION

DEFINITION
A hypergraph of text (HoT) is

$H = (V, E, T_V, T_E)$

- V set of nodes
- E set of hyperedges
- T_V set of node texts
- T_E set of hyperedge texts

02 / PROBLEM DEFINITION

Does this set of nodes form a valid hyperedge?

Given a candidate set of nodes $e \subseteq V$ with any textual description t, estimate the probability that it constitutes a valid hyperedge in H.

$$f_\theta(e, t) = \begin{cases} \geq \epsilon, & \text{if } e \in E \text{ and } t = t_e \in T_E \\ < \epsilon, & \text{otherwise} \end{cases}$$

CANDIDATE GROUP + DESCRIPTION → PROBABILITY

A SOLUTION CAN EXPLOIT

STRUCTURE · NODE TEXT · HYPEREDGE TEXT

Our three-branch architecture combines these complementary signals into richer representations for more accurate predictions.

03 / PROPOSED FRAMEWORK

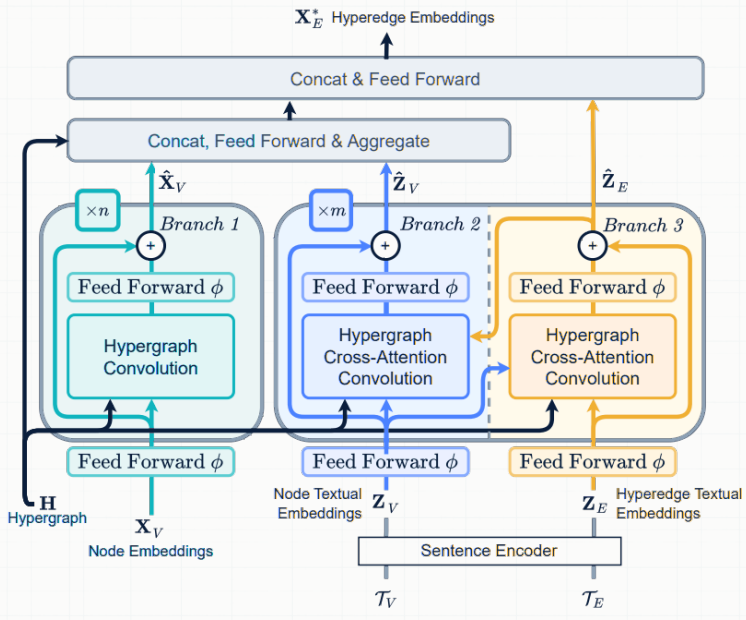
Three branches. One prediction.

SENTENCE ENCODER
Produces node-text and hyper edge-text embeddings.

BRANCH 1 / NODE STRUCTURE
Learns node embeddings using hypergraph convolution.

BRANCH 2 / NODE TEXT REFINEMENT
BRANCH 3 / HYPEREDGE TEXT REFINEMENT
Refines node-text and hyperedge-text representations with hypergraph cross-attention convolution.

AGGREGATION AND PREDICTION
Updated node representations are combined with hyperedge-text representations and used to predict the hyperedge score.



REPRESENTATION REFINING → CONCATENATE · AGGREGATE · CLASSIFY

04 / EXPERIMENTAL SETUP

TRAIN/VALIDATION/TEST SPLIT 80/10/10
INDEPENDENT TRIALS 10
BEST SETTING FOR OUR MODEL $n = 1, m = 3$
EARLY STOPPING PATIENCE $= 40$ epochs
NEGATIVE SAMPLING $\alpha = 0.8, \quad \beta = 3$

05 / DATASETS

Name	|V|	|E|	T_V	T_E
arXiv	16953 authors	54241 articles	✗	✓
Coursera	832 trainers	1104 courses	✗	✓
IMDb	20874 actors	8839 movies	✗	✓
Patent	17388 authors	5897 patents	✗	✗

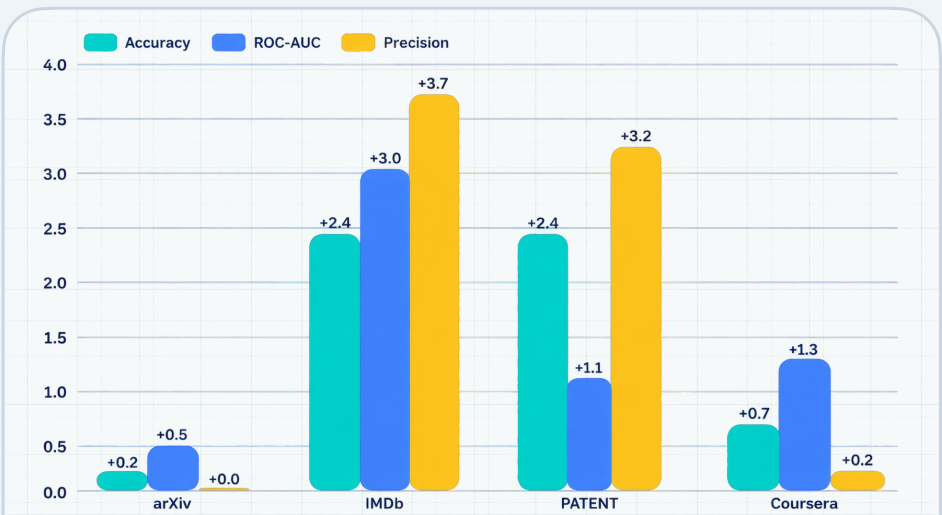
Missing texts were generated with GPT-4o-mini using prompts that exclude topological information.

06 / RESULTS

Topology provides the backbone. Semantics makes it stronger.

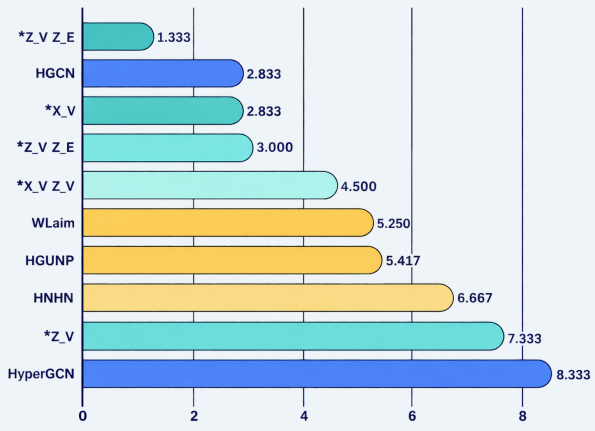
Text improves prediction over the four datasets.

Results are averaged over 10 runs on a machine equipped with an NVIDIA RTX 4000 GPU, an Intel Core i7 CPU, and 16 GB of RAM.



MEAN **pp** IMPROVEMENT OVER HGCN (PERCENTAGE-POINT)

*our model



MEAN RANK FOR EACH METRIC ACROSS DATASETS (Lower is better)

KEY TAKEAWAY

In our experiments combining structural connectivity with semantic information enhanced hyperlink prediction.

BEST OVERALL RANK

1.33

OUR FULL MODEL

HGCN BASELINE **2.833**

AVERAGE GAIN

+1.57pp

OVER HGCN

3 METRICS
4 DATASETS

PAPER

CODE

https://doi.org/10.1145/3799682.3839872

https://github.com/hypernetwork-research-group/hphot