



Structural Signal-based Inductive Reasoning in Knowledge Graph Foundation Model

노드 ID에 의존하지 않는 경로·관계 구조 신호를 KGFM의
추론 신호로 활용

Dong Hyun Lee(이동현)

Graph & Language Intelligence Laboratory
Department of Computer Science and Engineering
Konkuk University

2026.08.05

CONTENTS

1. Background
2. Entity-Inductive Reasoning
3. Structural Signal-based KGFM
4. LLM Integration Strategies
5. Research Direction



CONTENTS

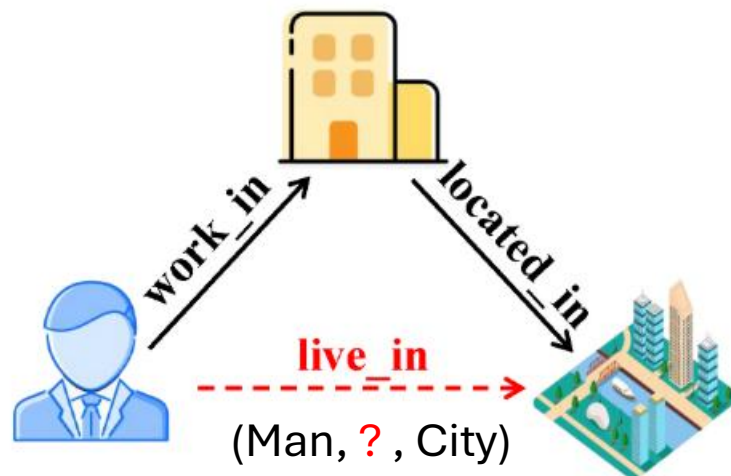
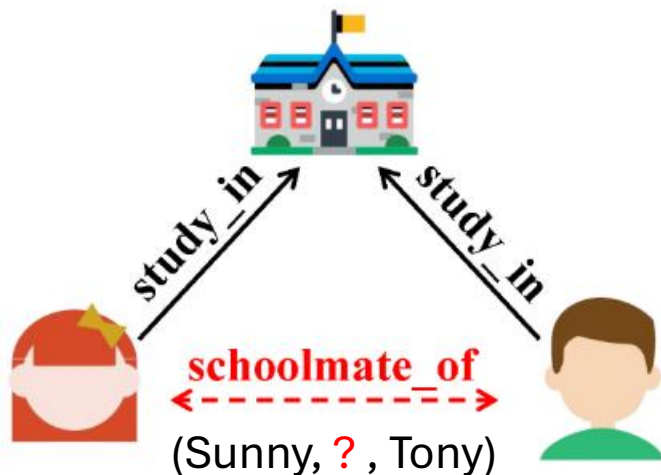
1. Background
2. Entity-Inductive Reasoning
3. Structural Signal-based KGFM
4. LLM Integration Strategies
5. Research Direction



What is Knowledge graph Foundation Model

Background : Knowledge Graph & Link Prediction

- Knowledge Graph는 world knowledge를 (head, relation, tail) triplet으로 표현
 - ✓ Entity는 node, relation type은 labeled edge에 대응
 - Ex) (Messi, plays_for, Barcelona)
 - ✓ Text corpus보다 구조적이며, relation path를 따라 logical reasoning이 가능
- Task : Link Prediction**
 - ✓ 현실의 KG는 본질적으로 incomplete하며, 이를 채우는 것이 목표
 - ✓ Query ($h\ r\ ?$) 또는 ($h\ ?\ t$)에 대해 missing fact를 추론
 - ✓ Recommender system, entity linking, KGQA, drug discovery 등에 활용



What is Knowledge graph Foundation Model

Background: Transductive KG Embedding

- 전통적 접근 : Transductive Setting
 - ✓ 각 entity와 relation에 고유한 learnable embedding을 할당하고 scoring function으로 plausibility를 계산
 - TransE, DistMult, ComplEx, RotatE 등
 - 학습된 embedding table 자체가 그 KG의 지식을 담는 구조
 - ✓ Transductive 가정: 학습 시점에 모든 entity·relation이 관측되어 있어야 함
 - 새로운 entity가 등장하면 대응 embedding이 없어 추론 자체가 불가능
 - 새로운 KG마다 재학습이 필요 → one graph, one model
 - embedding table이 graph-specific parameter인 근본적 제약 존재

What is Knowledge graph Foundation Model

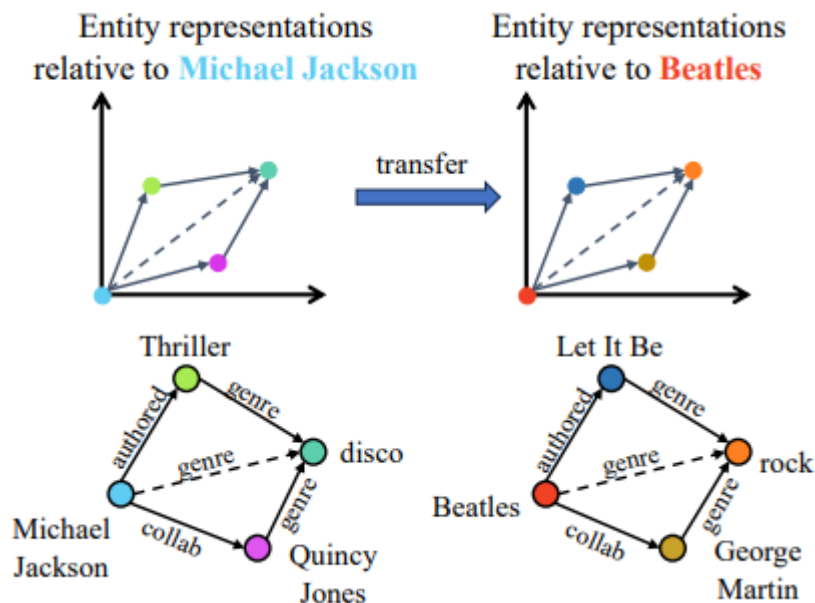
Background: Two Inductive Settings

- Training graph와 Test graph의 vocabulary 중첩 정도에 따른 inductive 상황
 - ✓ **Entity-Inductive** (Partially Inductive)
 - $\mathcal{E}_{\text{train}} \subset \mathcal{E}_{\text{test}}$, 단 $\mathcal{R}_{\text{train}} = \mathcal{R}_{\text{test}}$
 - 새로운 **entity**는 등장하지만 relation vocabulary는 고정
 - EX) 같은 도메인 KG에 신규 인물·상품이 추가되는 상황
 - ✓ **Fully-Inductive**
 - $\mathcal{E}_{\text{train}} \subset \mathcal{E}_{\text{test}}$, $\mathcal{R}_{\text{train}} \subset \mathcal{R}_{\text{test}}$
 - entity와 relation 모두 처음 보는 것 → 사실상 완전히 다른 domain의 KG
 - EX) Wikipedia KG로 학습 후 biomedical KG에서 추론

What is Knowledge graph Foundation Model

Background: Entity-Inductive Reasoning

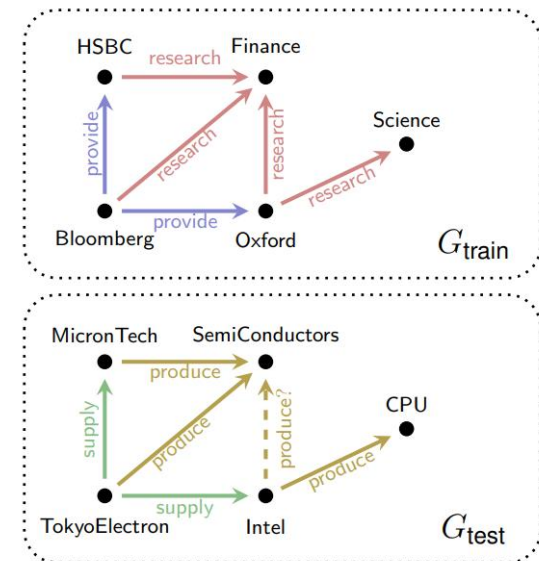
- 핵심 아이디어: entity를 "ID"가 아니라 "구조적 역할(structural role)"로 표현
 - ✓ Query의 head entity를 기준으로 삼아, 모든 node를 그에 대한 relative representation으로 계산
 - ✓ 학습되는 것은 특정 entity가 아니라 relational pattern
 - EX) $A \xrightarrow{\text{authored}} B \xrightarrow{\text{genre}} C$ 라는 패턴을 학습하면 Michael Jackson이든 Beatles 든 동일하게 적용됨
 - ✓ 그러나 relation은 여전히 학습된 embedding으로 표현됨



What is Knowledge graph Foundation Model

Background: Fully-Inductive Reasoning & KGFM

- Fully-Inductive를 풀기 위해서는 relation 역시 vocabulary와 무관하게 표현해야 함
 - ✓ relation들 사이의 상호작용 패턴을 이용해서, relation 표현 생성
 - ✓ 이 구조적 불변성을 학습하면 임의의 relational vocabulary를 가진 KG로 전이
- 이를 바탕으로 Knowledge Graph Foundation Model(KGFM) 패러다임이 등장 (2024)
 - ✓ one graph, one model → **one model, any graph**
 - ✓ Input feature 없이 구조 신호(**structural signal**)만으로 pre-train 후 **zero-shot inference**
 - ✓ 소수의 KG로 pre-train한 단일 모델이 50개 이상의 unseen KG에서 추론



What is Knowledge graph Foundation Model

Background: Why LLM Integration?

- 구조 신호만으로는 원리적으로 구분 불가능한 경우가 존재함
 - ✓ 구조는 같지만 의미가 다른 relation을 동일하게 표현하는 문제
 - ✓ EX) fatherOf vs sonOf — 구조적으로 대칭이지만 정반대의 예측 패턴
- Test-time relation vocabulary가 희소한 setting에서 구조 기반은 성능이 급격히 붕괴
- 실세계 KG는 entity·relation에 자연어 텍스트를 거의 항상 동반하지만, 구조 기반 KGFM은 이를 활용하지 않음



CONTENTS

1. Background
2. Entity-Inductive Reasoning
 - [2022][ICLR][NBFNet]
3. Structural Signal-based KGFM
4. LLM Integration Strategies
5. Research Direction

[2022][NIPS][NBFNet]

Neural Bellman-Ford Networks: A General Graph Neural Network Framework for Link Prediction

Zhaocheng Zhu^{1,2}, Zuobai Zhang^{1,2}, Louis-Pascal Xhonneux^{1,2}, Jian Tang^{1,3,4}

Mila - Québec AI Institute¹, Université de Montréal²

HEC Montréal³, CIFAR AI Chair⁴

{zhaocheng.zhu, zuobai.zhang, louis-pascal.xhonneux}@mila.quebec
jian.tang@hec.ca

Entity-Inductive Reasoning

[2022][NIPS][NBFNet] : Motivation

- Link prediction의 기존 패러다임은 각각 명확한 trade-off
 - ✓ Embedding-based (TransE, DistMult, RotatE)
 - local subgraph를 명시적으로 encoding하지 않으며 transductive
 - ✓ GNN-based (VGAE, RGCN / SEAL, GraIL)
 - Auto-encoder 계열은 표현력은 높지만 transductive
 - Subgraph 계열은 inductive, link마다 subgraph를 만들어야 해 scalability 붕괴
- 연구 질문 : Inductive generalization · Interpretability · High capacity · Scalability를 동시에 만족할 수 있을까?

Method	Inductive ³	Interpretable	Learned Representation	Time Complexity	Wall Time
VGAE [32] / RGCN [48]			✓	$O(d)$	18 secs
NeuralLP [69] / DRUM [46]	✓	✓		$O\left(\frac{ \mathcal{E} d}{ \mathcal{V} } + d^2\right)$	2.1 mins
SEAL [73] / GraIL [55]	✓		✓	$O(\mathcal{E} d^2)$	≈1 month
NBFNet	✓	✓	✓	$O\left(\frac{ \mathcal{E} d}{ \mathcal{V} } + d^2\right)$	4.0 mins

Entity-Inductive Reasoning

[2022][NIPS][NBFNet] : Method

- Key Idea

- ✓ node pair의 표현을 "두 노드 사이 모든 path 표현의 일반화된 합"으로 정의
- ✓ Query relation q 에 대한 pair representation

$$h_q(u, v) = h_q(P_1) \oplus h_q(P_2) \oplus \dots \oplus h_q(P_{|\mathcal{P}_{uv}|}) \Big|_{P_i \in \mathcal{P}_{uv}} \triangleq \bigoplus_{P \in \mathcal{P}_{uv}} h_q(P)$$

$$h_q(P = (e_1, e_2, \dots, e_{|P|})) = w_q(e_1) \otimes w_q(e_2) \otimes \dots \otimes w_q(e_{|P|}) \triangleq \bigotimes_{i=1}^{|P|} w_q(e_i)$$

(u,v)의 표현 : u에서 v에 이르는 모든 경로들의 합

경로의 표현 : 경로를 구성하는 모든 Edge의 곱

Entity-Inductive Reasoning

[2022][NIPS][NBFNet] : Method

- Generalized Bellman-Ford Algorithm
 - ✓ 문제: Path 수는 길이에 따라 지수적으로 증가 $\rightarrow$ 직접 열거는 불가능
 - ✓ Berman Fold : u 에서 v 로 가는 모든 경로는, 반드시 v 의 이웃 중 하나(x)를 마지막에서 두 번째로 거친다.

$$\mathcal{P}_{uv} = \bigcup_{(x,r,v) \in \mathcal{E}(v)} \left\{ (u \rightarrow x \text{ 경로}) + \text{마지막 edge } (x, r, v) \right\}$$



재귀적으로 정의

$$\mathbf{h}^{(t)}(v) = \bigoplus_{(x,r,v)} \mathbf{h}^{(t-1)}(x) \otimes \mathbf{w}_q(x, r, v)$$

Entity-Inductive Reasoning

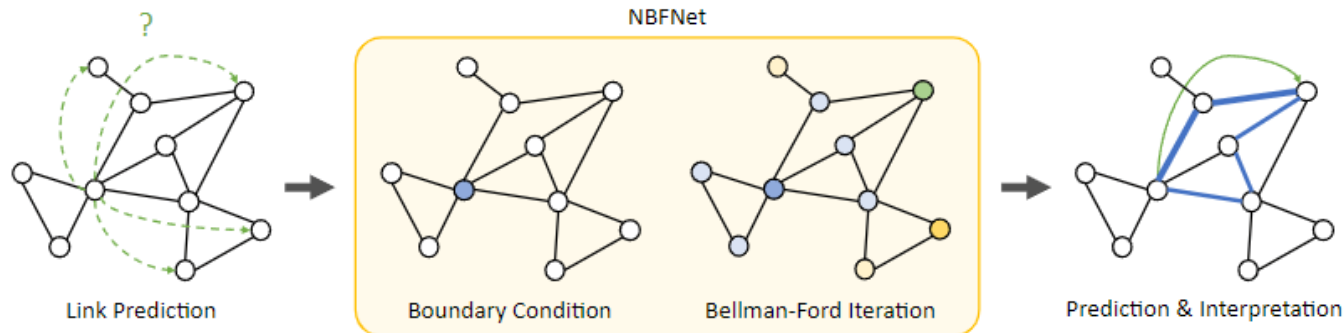
[2022][NIPS][NBFNet] : Method

최종 형태

$$\mathbf{h}_v^{(0)} \leftarrow \text{INDICATOR}(u, v, q)$$

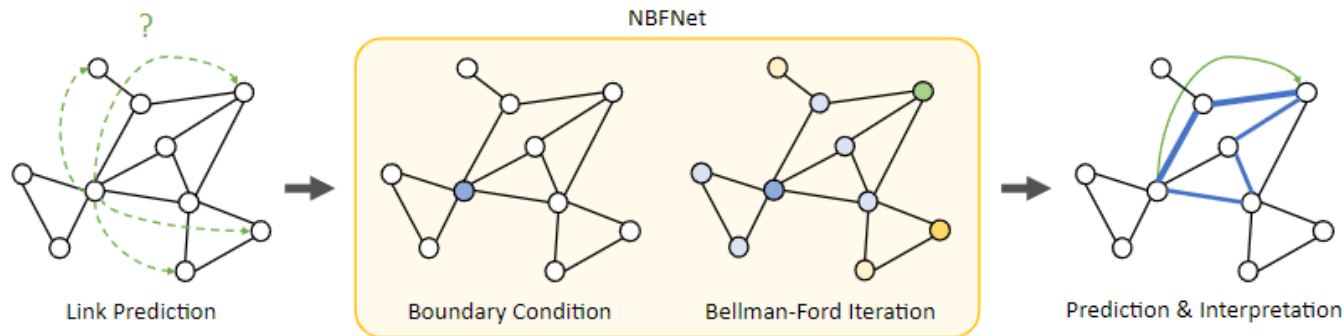
$$\mathbf{h}_v^{(t)} \leftarrow \text{AGGREGATE} \left(\left\{ \text{MESSAGE} \left(\mathbf{h}_x^{(t-1)}, \mathbf{w}_q(x, r, v) \right) \mid (x, r, v) \in \mathcal{E}(v) \right\} \cup \left\{ \mathbf{h}_v^{(0)} \right\} \right)$$

- ✓ INDICATOR : Query Entity는 1로 두고, 나머지는 0으로 둔다
- ✓ 즉, 우리가 아는 MPNN 에서 Feature를 삭제하고 쿼리만 1, 나머지는 0으로 둔 것
 - Entity의 Feature를 사용하지 않음으로써, Entity-Inductive 상황에 대한 일반화



Entity-Inductive Reasoning

[2022][NIPS][NBFNet] : Method



- 전체 파이프라인

1. Link Prediction ($h, r, ?$) 입력
2. Query Entity (h)를 1로 초기화, 나머지 Entity 0으로 초기화
3. Message passing Neural Network 진행
4. 각각의 Entity Embedding $\rightarrow$ MLP $\rightarrow$ 최종 Score

CONTENTS



1. Background
2. Entity-Inductive Reasoning
3. Structural Signal-based KGFM
 - [2024][ICLR][ULTRA]
 - [2025][ICML][MOTIF]
 - [2026][ICLR][FLOCK]
4. LLM Integration Strategies
5. Research Direction

[2024][ICLR][ULTRA]

TOWARDS FOUNDATION MODELS FOR KNOWLEDGE GRAPH REASONING

Mikhail Galkin^{1*}, Xinyu Yuan^{2,3}, Hesham Mostafa¹, Jian Tang^{2,4}, Zhaocheng Zhu^{2,3}

¹Intel AI Lab, ²Mila ³University of Montréal ⁴HEC Montréal & CIFAR AI Chair

Structural Signal-based KGFM

[2024][ICLR][ULTRA] : MOTIVATION

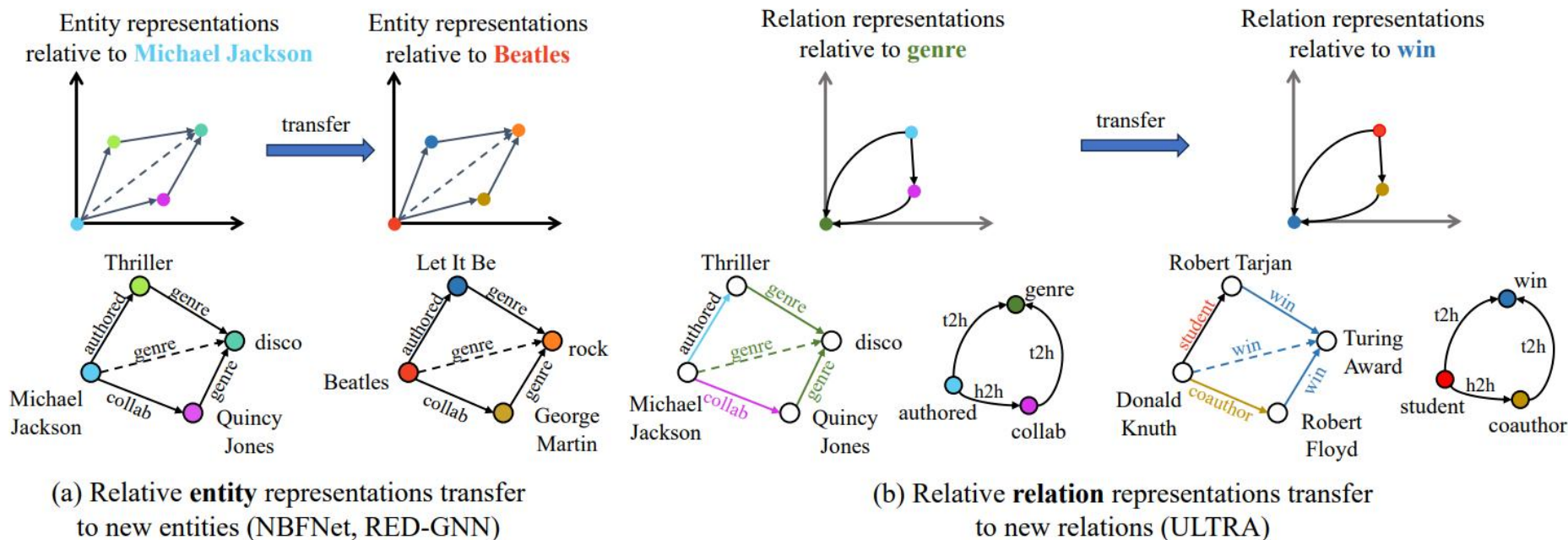
- Foundation model
 - ✓ Foundation model은 "도메인 불변의 어휘(vocabulary)"가 있기에 성립한다
 - ✓ 그러나 KG는 graph마다 entity-relation vocabulary가 다르고, 서로 겹치지 않는다
 - ✓ 대부분의 inductive 기법은 **entity 표현을 relation의 함수로 구성**하므로, relation vocabulary가 고정
- => 핵심 연구 질문: 임의의 entity-relation vocabulary를 가로지르는 invariance는 무엇인가?

Structural Signal-based KGFM

[2024][ICLR][ULTRA] : MOTIVATION

- ULTRA의 두 가지 핵심 관찰
 - ✓ Relation은 dataset마다 달라도, relation 사이의 상호작용은 유사하고 transferable
 - ✓ 초기 relation 표현을 그 상호작용에 조건화하면, 어떤 input feature도 필요 없다

⇒ 목표: entity:relation embedding을 전혀 학습하지 않고, 어떤 KG에도 zero-shot 적용 가능한 표현 학습

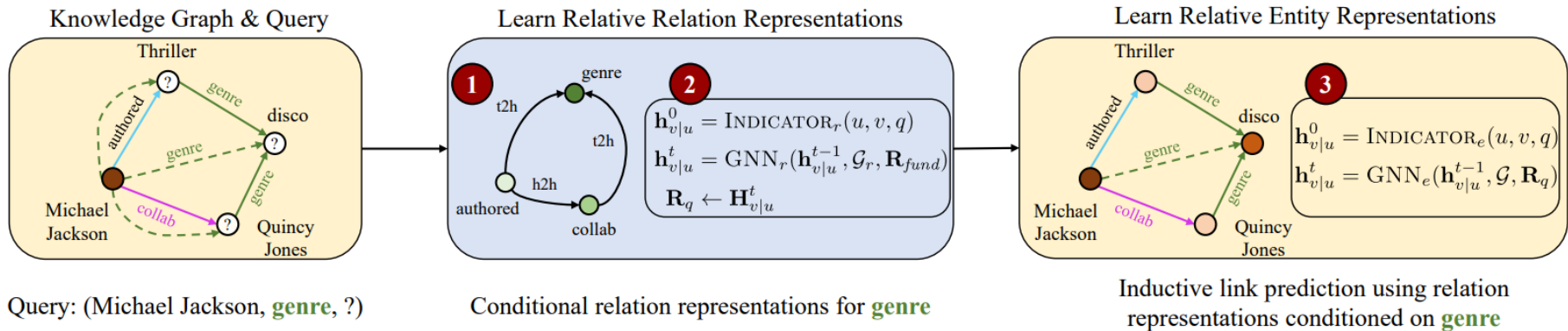


Structural Signal-based KGFM

[2024][ICLR][ULTRA] : METHOD

- ULTRA의 2-STAGE FRAMEWORK

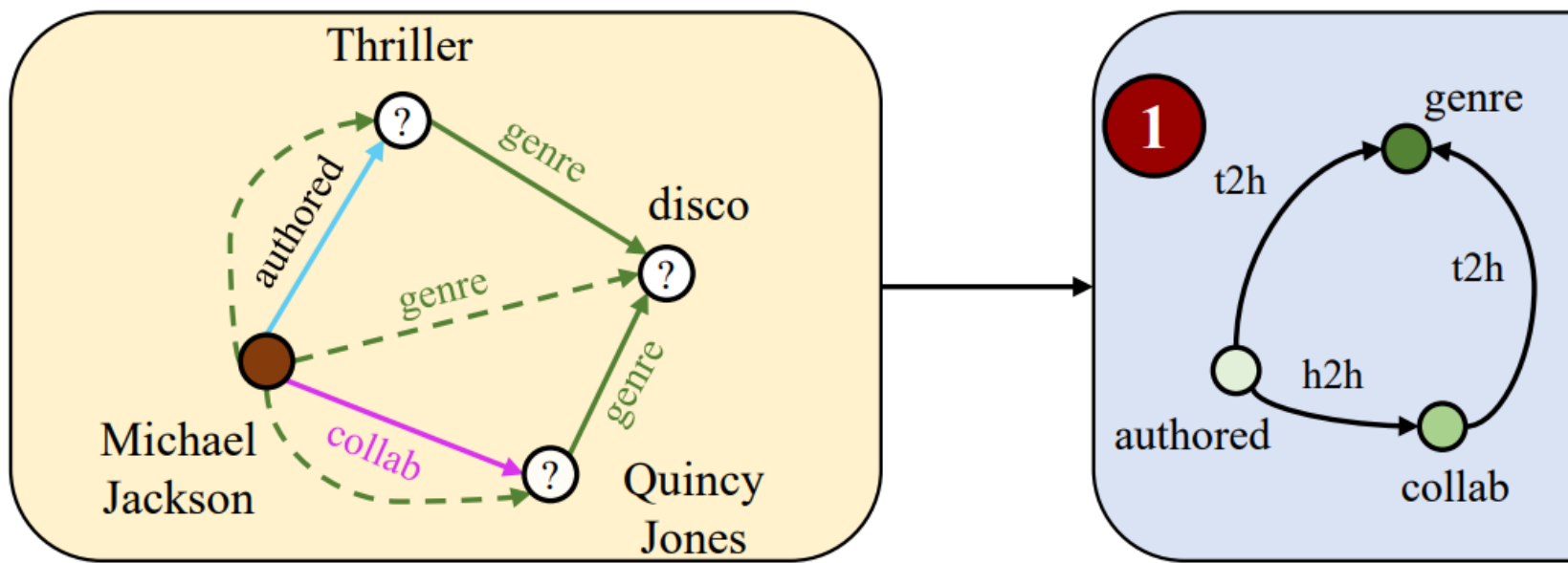
1. Original Graph에서 Relation간의 상호작용으로 Relation Graph 구성
2. STAGE1 : Relation Graph에서 NBFNet으로 구조적 Relation 표현 생성
3. STAGE2 : 생성된 Relation Embedding을 사용하여 원래 그래프에서 NBFNet



Structural Signal-based KGFM

[2024][ICLR][ULTRA] : METHOD

- LIFTING Relation Graph
 - **t2h**: $r_1(w\ v) \wedge r_2(v\ w)$ — r_1 의 tail이 r_2 의 head
 - **h2h**: $r_1(v\ u) \wedge r_2(v\ w)$ — head를 공유
 - **h2t**: $r_1(v\ u) \wedge r_2(w\ v)$
 - **t2t**: $r_1(w\ v) \wedge r_2(w\ v)$ — tail을 공유



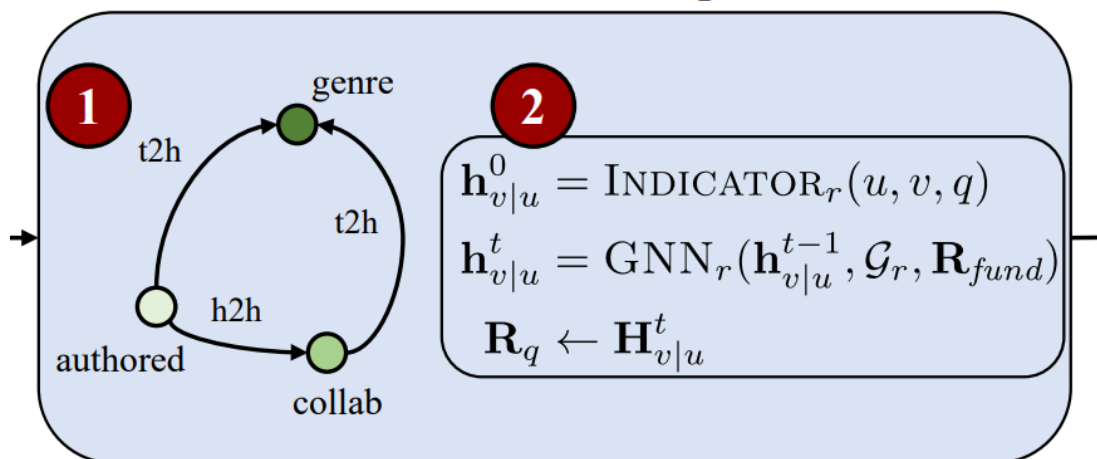
=> Relation간의 상호작용을 포착하기 위함!

Structural Signal-based KGFM

[2024][ICLR][ULTRA] : METHOD

- Stage1 : Relation Graph 위에서 NBFNet을 진행
 - ✓ NBFNet의 labeling trick을 **relation 층위로 그대로 끌어올림**
 - ✓ Query Relation = 1, 나머지 Relation = 0
 - ✓ 결국 학습하는 것은 t2h 같은 edge의 Representation / MPNN Linear layer

Learn Relative Relation Representations

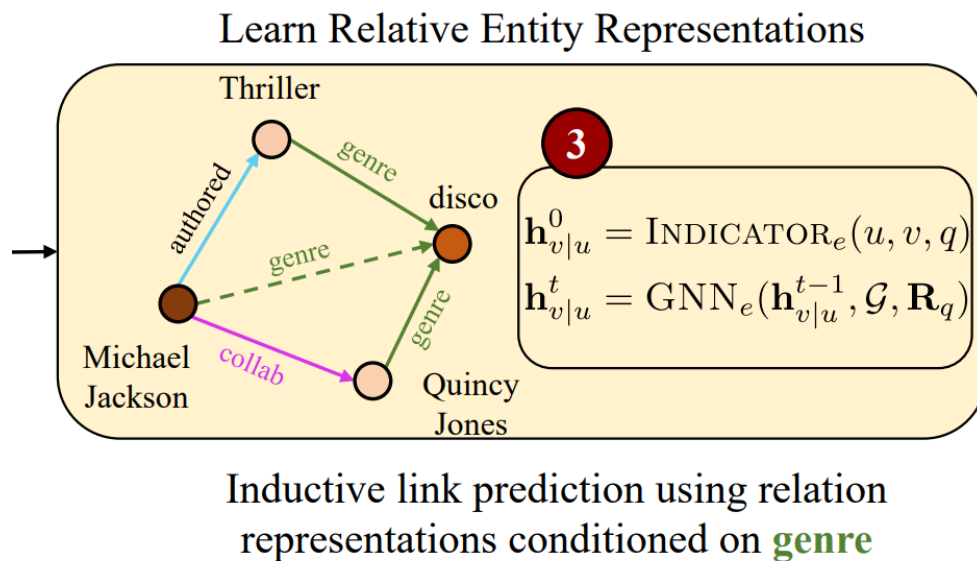


=> Feature 정보를 포함하지 않는 **구조적 Relation 표현 획득**

Structural Signal-based KGFM

[2024][ICLR][ULTRA] : METHOD

- Stage2 : Original Graph 위에서 NBFNet을 진행
 - ✓ Relation 표현 = STAGE1에서 가져온 구조적 표현
 - ✓ NBFNet과 똑같이, Query Entity를 1로 두고, 나머지 Entity는 0으로 둔다
 - ✓ 학습하는 것은 Entity Layer 뿐



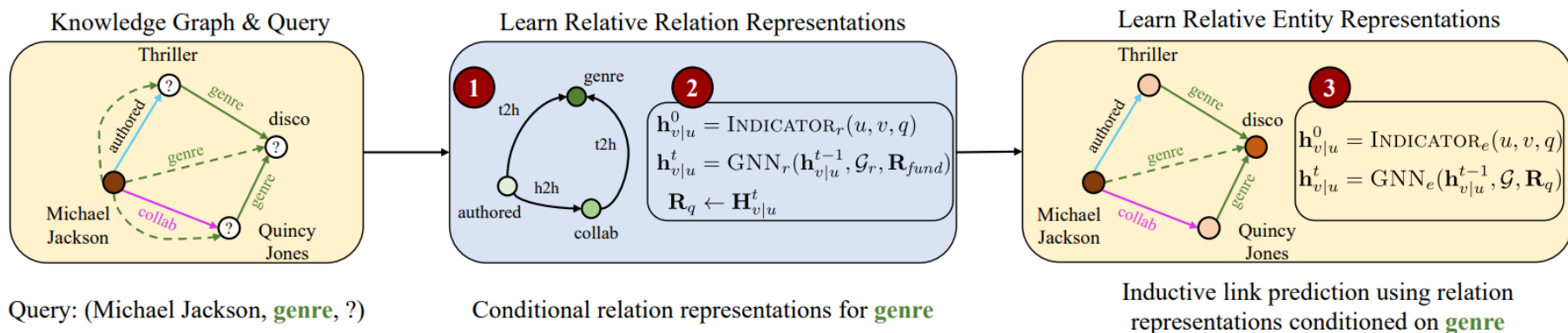
=> Feature 정보를 포함하지 않는 구조적 Relation 표현 획득

Structural Signal-based KGFM

[2024][ICLR][ULTRA] : METHOD

- ULTRA의 2-STAGE FRAMEWORK

1. Original Graph에서 Relation간의 상호작용으로 Relation Graph 구성
2. STAGE1 : Relation Graph에서 NBFNet으로 구조적 Relation 표현 생성
3. STAGE2 : 생성된 Relation Embedding을 사용하여 원래 그래프에서 NBFNet



2-STAGE 구조를 통해, Relation / Entity의 Feature를 전부 사용하지 않음
=> 학습 시점에 못 본 Relation/Entity에 대해서 일반화가 가능

Structural Signal-based KGFM

[2024][ICLR][ULTRA] : EXPERIMENT

Model	Inductive (e) + (e, r) (27 graphs)		Transductive e (13 graphs)		Total Avg (40 graphs)		Pretraining (3 graphs)		Inductive (e) + (e, r) (8 graphs)
	MRR	H@10	MRR	H@10	MRR	H@10	MRR	H@10	Hits@10 (50 negs)
Supervised SOTA	0.342	0.482	0.348	0.494	0.344	0.486	0.439	0.585	0.731
ULTRA 0-shot	0.435	0.603	0.312	0.458	0.395	0.556	-	-	0.859
ULTRA fine-tuned	0.443	0.615	0.379	0.543	0.422	0.592	0.407	0.568	0.896

- ULTRA의 Pretraining
 - ✓ FB15k-237 + WN18RR + CoDEx-Medium (Freebase / WordNet / Wikidata 계열)
 - ✓ 학습되는 것: 4개 fundamental interaction embedding + UPDATE linear layer **뿐**
 - Entity embedding도, relation embedding도 학습하지 않음
- ULTRA의 Inference
 - ✓ 57개 KG(1K–120K nodes)에서 검증
 - ✓ **Zero-shot** 평균 MRR 0.395 vs graph별 **supervised SOTA** 0.344 → 학습 없이 평균 **15% 상회**

[2025][ICML][MOTIF]

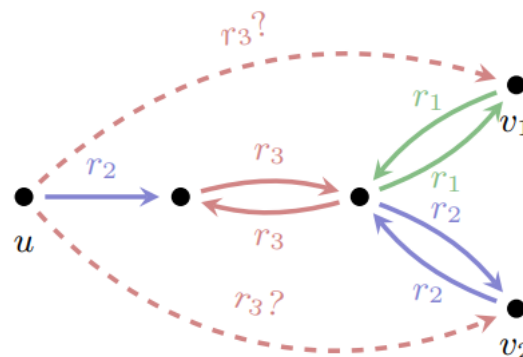
How Expressive are Knowledge Graph Foundation Models?

**Xingyue Huang¹ Pablo Barceló^{2,3,4} Michael M. Bronstein^{1,5} İsmail İlkan Ceylan¹ Mikhail Galkin⁶
Juan L Reutter^{2,3,4} Miguel Romero Orth^{2,4}**

Structural Signal-based KGFM

[2025][ICML][MOTIF] : MOTIVATION

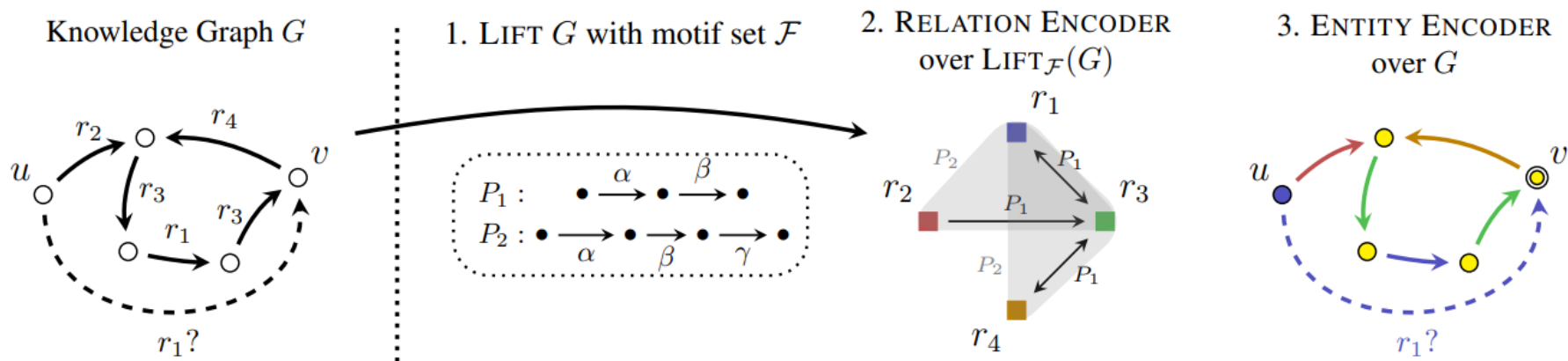
- ULTRA가 KGFM의 표준이 되었지만, "얼마나 표현력이 있는가"는 한 번도 분석된 적이 없음
 - ✓ ULTRA가 relation 사이에서 보는 정보는 사실 매우 제한적임
 - ✓ ULTRA의 네 interaction(h2h, h2t, t2h, t2t)은 relation 두 개짜리 2-path 패턴
 - ✓ 세 개 이상의 relation이 함께 만들어내는 higher-order 상호작용은 표현 불가
- 그 결과, 구조적으로 구분되어야 할 relation들이 같은 표현으로 붕괴함
 - ✓ Relation 표현이 뭉개지면 entity encoder가 이웃 node를 구분할 수 없게 되어 link prediction 자체가 실패



Structural Signal-based KGFM

[2025][ICML][MOTIF] : METHOD

- 핵심 아이디어: "relation graph"를 "relational hypergraph"로 일반화
 - ✓ ULTRA의 4개 interaction을 특수 케이스로 포함하는 general motif framework
 - ✓ 파이프라인은 ULTRA와 동일한 3단계, ①만 확장됨
 - ✓ 기존의 Relation Graph 구축 -> Relation Hypergraph 구축
 - Motif = relation k 개로 이루어진 작은 패턴. 그래프 안에 그 패턴이 실제로 존재하면, 그 k 개 relation을 한꺼번에 묶는 edge를 긋는다.

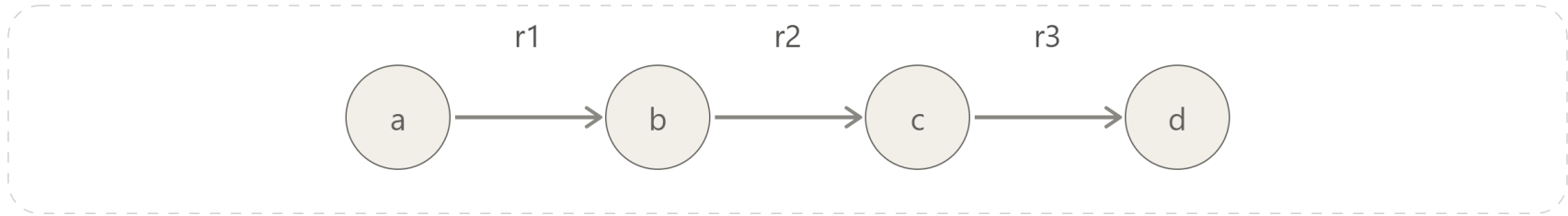


=> "ULTRA는 relation을 둘씩만 봤다. MOTIF는 셋 이상을 한꺼번에 본다"

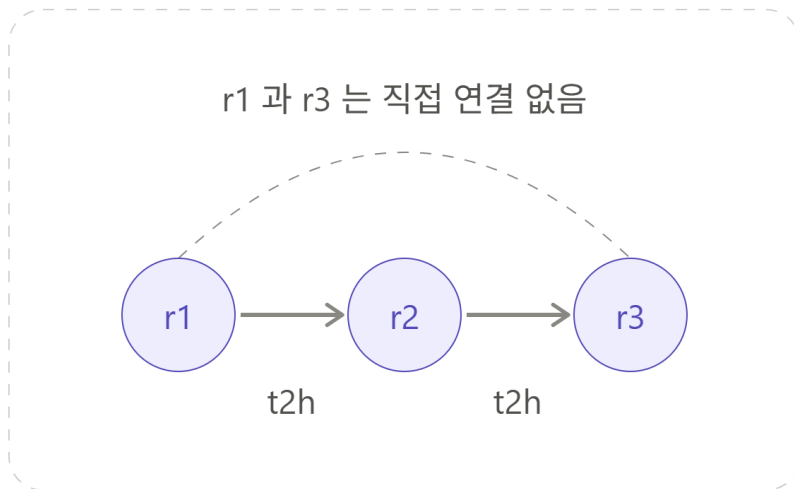
Structural Signal-based KGFM

[2025][ICML][MOTIF] : METHOD

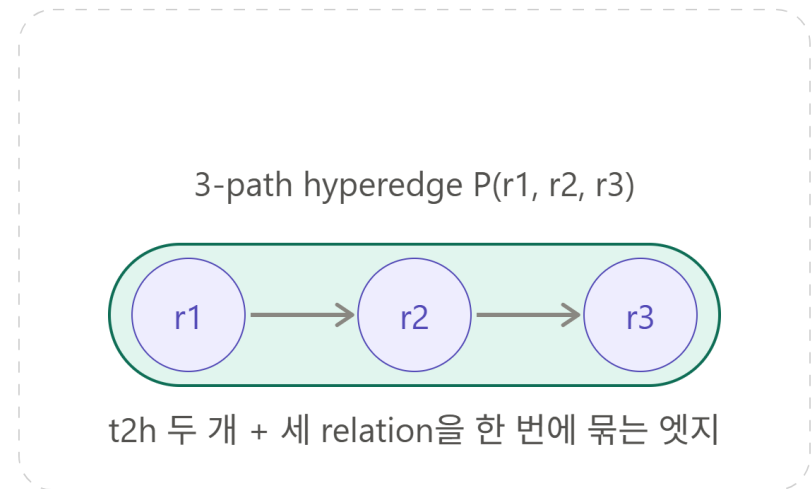
입력 knowledge graph



① ULTRA — relation graph



② MOTIF — relational hypergraph



=> Relation Graph를 구성하는 Edge의 종류가 Relation 그래프의 표현력을 결정한다

[2026][ICLR][FLOCK]

FLOCK: A KNOWLEDGE GRAPH FOUNDATION MODEL VIA LEARNING ON RANDOM WALKS

Jinwoo Kim^{1*} Xingyue Huang^{2*} Krzysztof Olejniczak² Kyungbin Min¹
Michael Bronstein^{2,4} Seunghoon Hong¹ İsmail İlkan Ceylan^{3,4,2}
¹KAIST ²University of Oxford ³TU Wien ⁴AITHYRA

Structural Signal-based KGFM

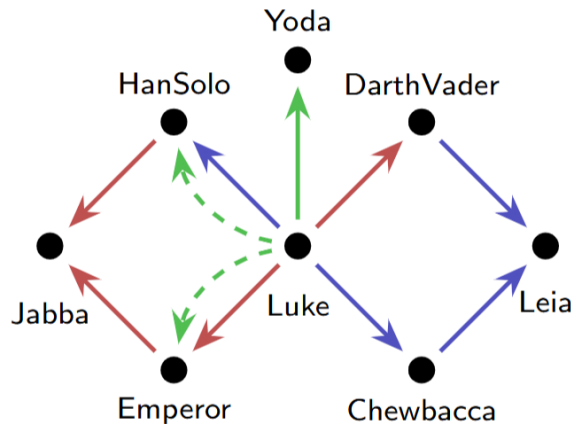
[2026][ICLR][FLOCK] : MOTIVATION

- The Two-Stage Paradigm
 - ✓ ULTRA – BASED 방법은 2-STAGE 방법을 통해 구성됨
 - ✓ 즉, relation을 먼저 인코딩하고 그 다음 Entity를 인코딩
 - ✓ 이로 인한 정보 손실 발생
 - Entity의 구조적 정보나, 빈도수 등의 정보가 손실
 - Relation 표현의 왜곡으로 이어짐

Structural Signal-based KGFM

[2026][ICLR][FLOCK] : MOTIVATION

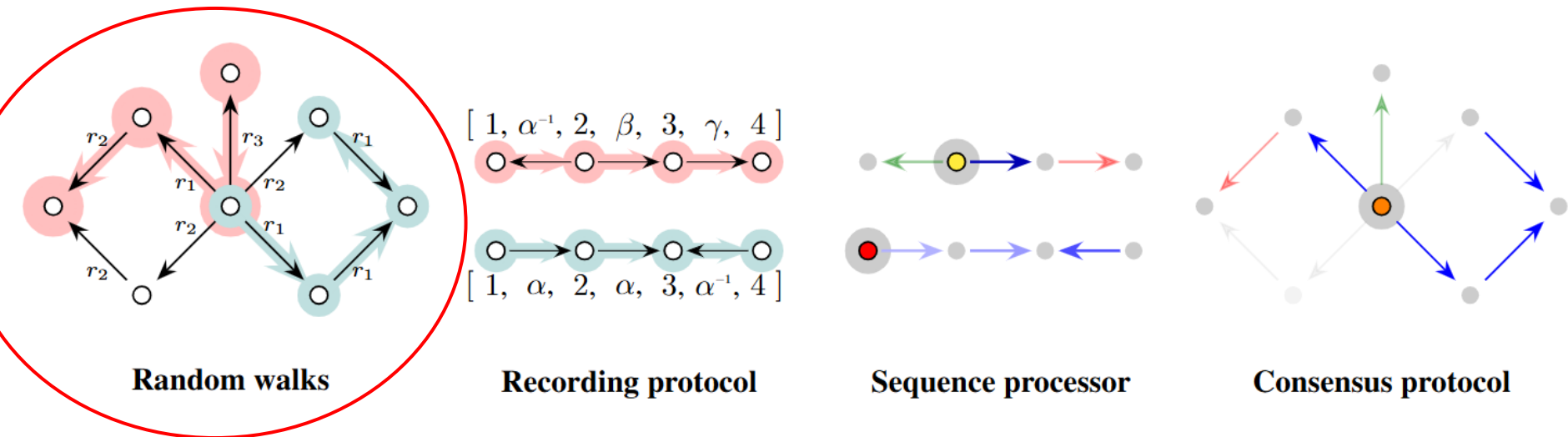
- Expressivity Ceiling of Deterministic Equivariance
 - ✓ 기존 KGFM의 요구 조건 : "구조적으로 isomorphic한 relation은 반드시 동일한 표현을 가져야 한다"
 - ✓ 이 제약이 KG 간 일반화의 원천이지만, 동시에 표현력의 천장이 됨
 - ✓ 결과: 구조는 같지만 의미가 다른 relation을 원리적으로 구분 불가
- SOLUTION : 제약을 "확률 수준"으로 완화한다 Probabilistic Equivariance
 - ✓ message passing 없이, Random walk를 사용
 - ✓ 즉, Random Walk의 랜덤성으로 인해 같은 구조가 한 번의 Forward에서 다른 표현이 가능함



Structural Signal-based KGFM

[2026][ICLR][FLOCK] : METHOD

- ① Random Walk Algorithm
 - ✓ Diverse starting locations
 - query ($h, r, ?$)에 대해 총 $3n$ 개 walk
 - 쿼리 노드에서 출발 n 개, 무작위 Edge에서 출발 n 개, 무작위 노드에서 출발 n 개
 - ✓ Uniform non-backtracking walk
 - 직전 노드로 되돌아가지 않아 먼 영역까지 빠르게 탐색



Structural Signal-based KGFM

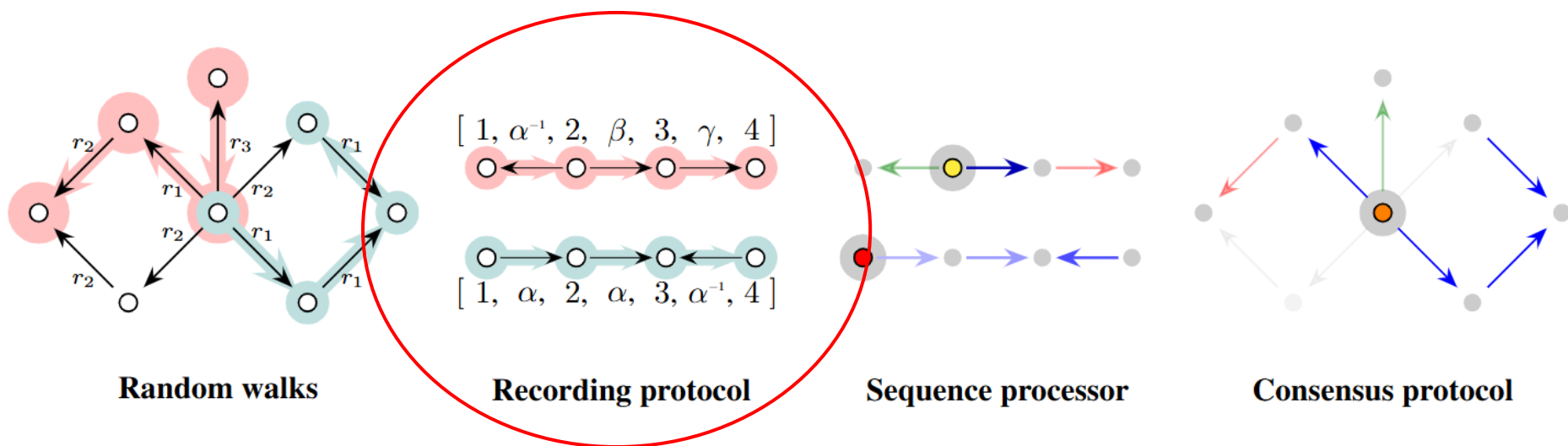
[2026][ICLR][FLOCK] : METHOD

- ② Recording Protocol

- ✓ Walk에는 KG마다 고유한 node·relation ID가 그대로 노출되어 전이 불가
- ✓ 해결 : Node-relation anonymization

$$\eta = v_0 \xrightarrow{r_1} v_1 \xrightarrow{r_2} v_2 \xrightarrow{r_1^{-1}} v_0 \quad \mapsto \quad 1 \xrightarrow{\alpha} 2 \xrightarrow{\beta} 3 \xrightarrow{\alpha^{-1}} 1,$$

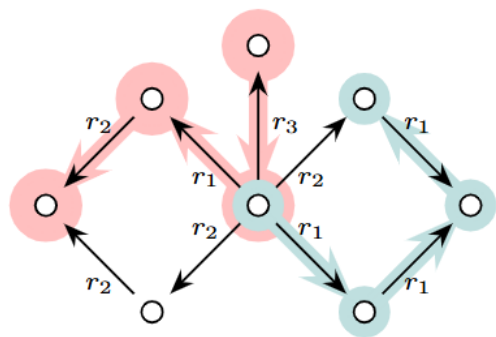
- ✓ node와 relation에 별도 namespace를 두고 등장 순서대로 이름 부여
-> 구조적 역할만 남고 identity는 사라짐



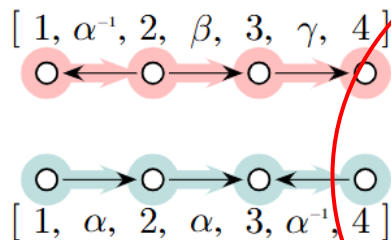
Structural Signal-based KGFM

[2026][ICLR][FLOCK] : METHOD

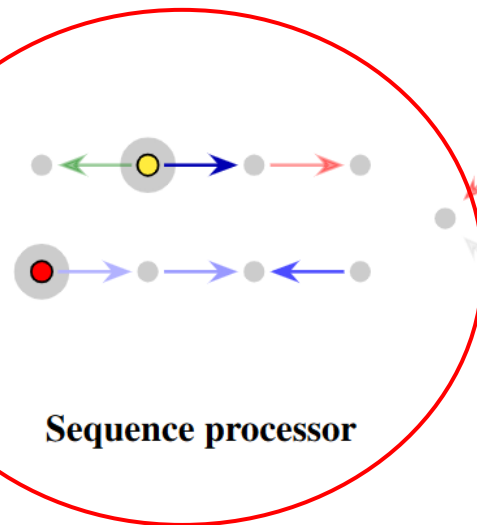
- ③ Sequence Processor — 각 walk를 독립적으로 인코딩
 - ✓ **Bidirectional GRU** (1 layer, hidden dim 64) 채택
 - ✓ 여기서 entity와 relation이 하나의 시퀀스 안에서 함께 인코딩됨



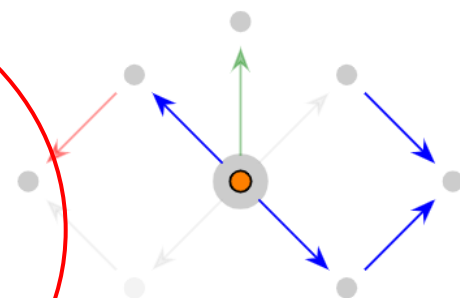
Random walks



Recording protocol



Sequence processor

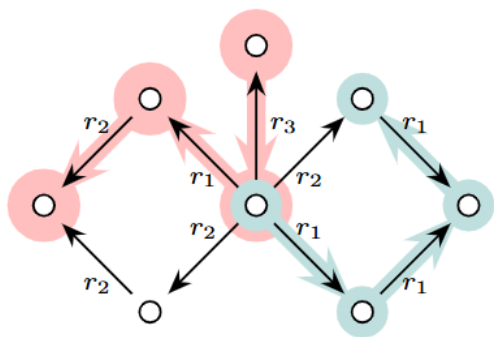


Consensus protocol

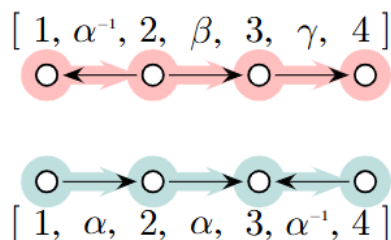
Structural Signal-based KGFM

[2026][ICLR][FLOCK] : METHOD

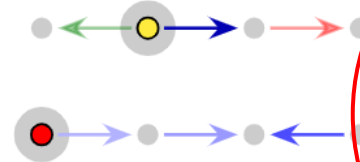
- ④ Consensus Protocol — walk들의 제안을 모아 state를 갱신
 - ✓ 각 walk는 자신이 방문한 node-relation에 대해 Update Protocol을 내놓는다
 - ✓ 이를 multi-head softmax-weighted average로 통합
 - ✓ 각 Walk에 해당하는 node-relation Embedding을 갱신
 - ✓ ① ~ ④ 과정을 반복하여 최종 임베딩을 획득



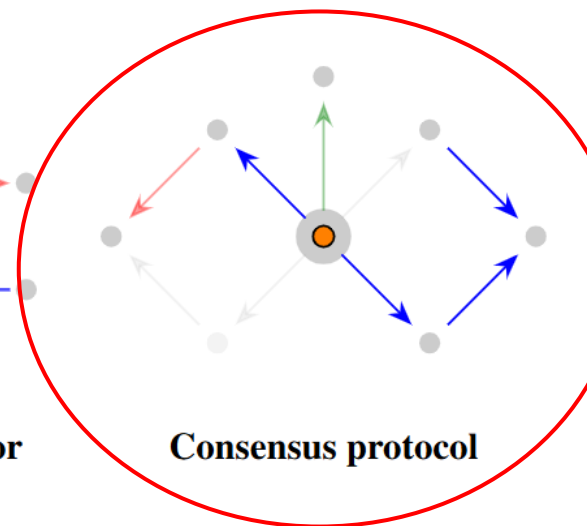
Random walks



Recording protocol



Sequence processor



Consensus protocol

CONTENTS

1. Background
2. Entity-Inductive Reasoning
3. Structural Signal-based KGFM
4. LLM Integration Strategies
 - [2025][EMNLP][SEMMA]
 - [2026][ICLR][KRLM]
5. Research Direction



[2024][ACL-F][PROLINK]

LLM as Prompter: Low-resource Inductive Reasoning on Arbitrary Knowledge Graphs

Kai Wang[†], Yuwei Xu[†], Zhiyong Wu[‡], Siqiang Luo^{†*}

[†]Nanyang Technological University [‡]Shanghai Artificial Intelligence Laboratory

[†]kai_wang@ntu.edu.sg, S220060@e.ntu.edu.sg, siqiang.luo@ntu.edu.sg

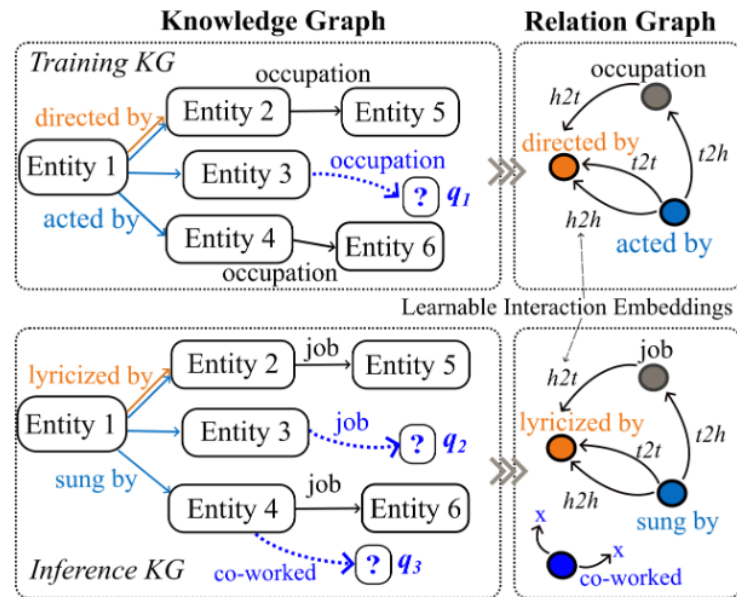
[‡]wuzhiyong@pjlab.org.cn

LLM Integration Strategies

[2024][ACL-F][PROLINK]

- KGFM은 relation graph의 구조 신호에 전적으로 의존하는데, 그 신호가 희박한 상황이 현실에 흔하다
 - ✓ **Low-resource inductive reasoning** — inference KG에서 query relation의 support triple이 거의 없거나 아예 없는 경우
 - ✓ EX) 신흥 도메인, 전문 도메인, 새로 추가된 relation 등 long-tail 분포

- 문제 정의: K-shot Inductive Reasoning
 - ✓ Fully-inductive 설정 위 새 relation r_q 에 대해 support triple이 단 K 개만 주어짐



LLM Integration Strategies

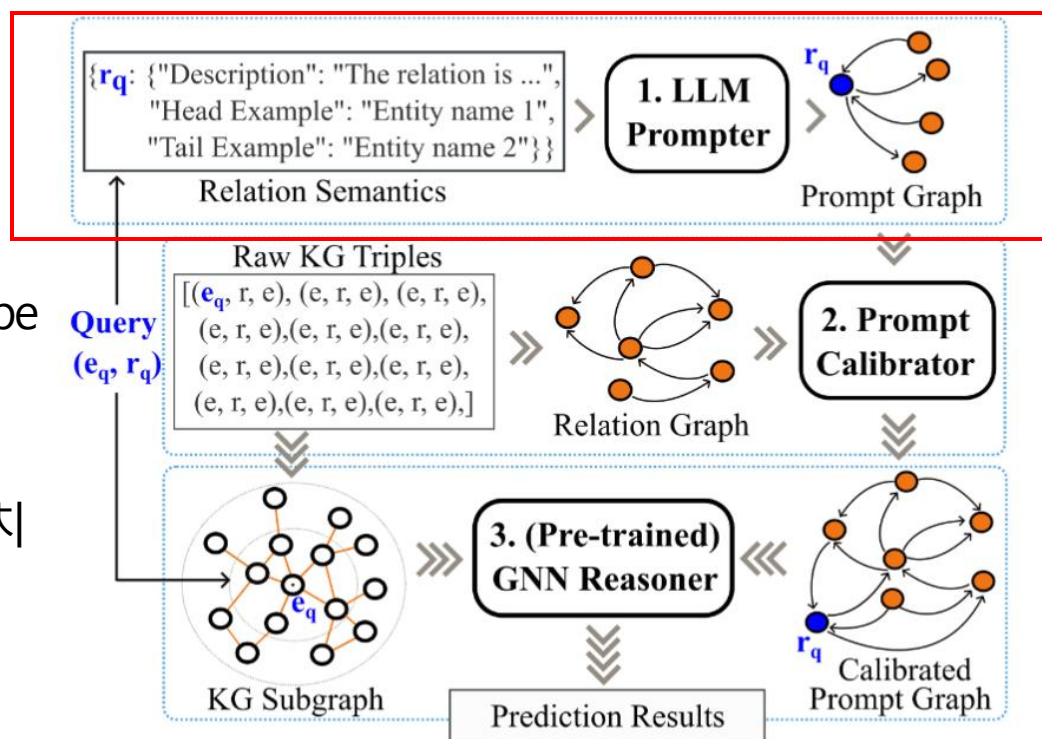
[2024][ACL-F][PROLINK]

- GNN Reasoner — 기본 골격은 ULTRA와 동일
 - ✓ 여기에 low-resource 대응을 위한 두 가지 강화 기법을 추가

- ① LLM Prompter

1. LLM을 통해, relation 하나당 단 한 번, head/tail 쪽의 후보 entity type을 질문한다

2. 두 relation의 head 쪽 type이 겹치면 h2h 엮지로 간주
(나머지 interaction도 동일 규칙)



LLM Integration Strategies

[2024][ACL-F][PROLINK]

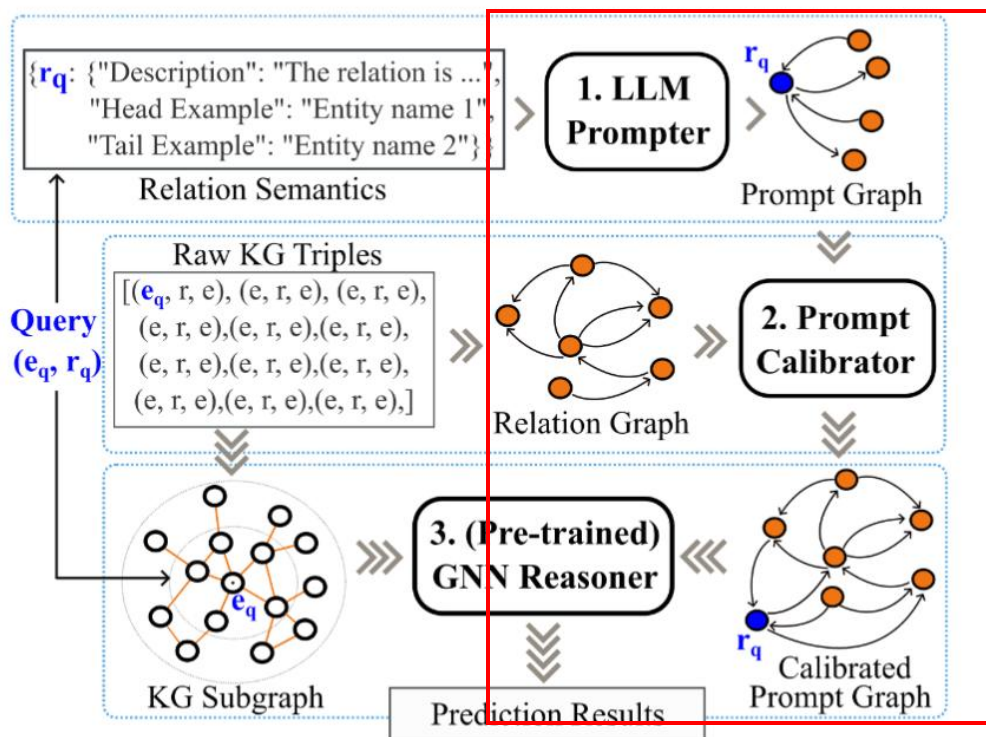
- GNN Reasoner — 기본 골격은 ULTRA와 동일
 - ✓ 여기에 low-resource 대응을 위한 두 가지 강화 기법을 추가

- ② Prompt Calibrator

- ✓ LLM이 만든 엣지의 노이즈 제거

- ✓ $\mathcal{G}_p$ 와 $\mathcal{G}_r$ 의 불일치 엣지 $\mathcal{T}_m$ 을 세어, 임계값 β 이상 충돌하는 relation은 r_q 와의 연결을 제거

EX) directed by와 lyricized by는
tail이 모두 people이지만
실제 entity가 겹치지 않음



[2025][EMNLP][SEMMA]

SEMMA: A Semantic Aware Knowledge Graph Foundation Model

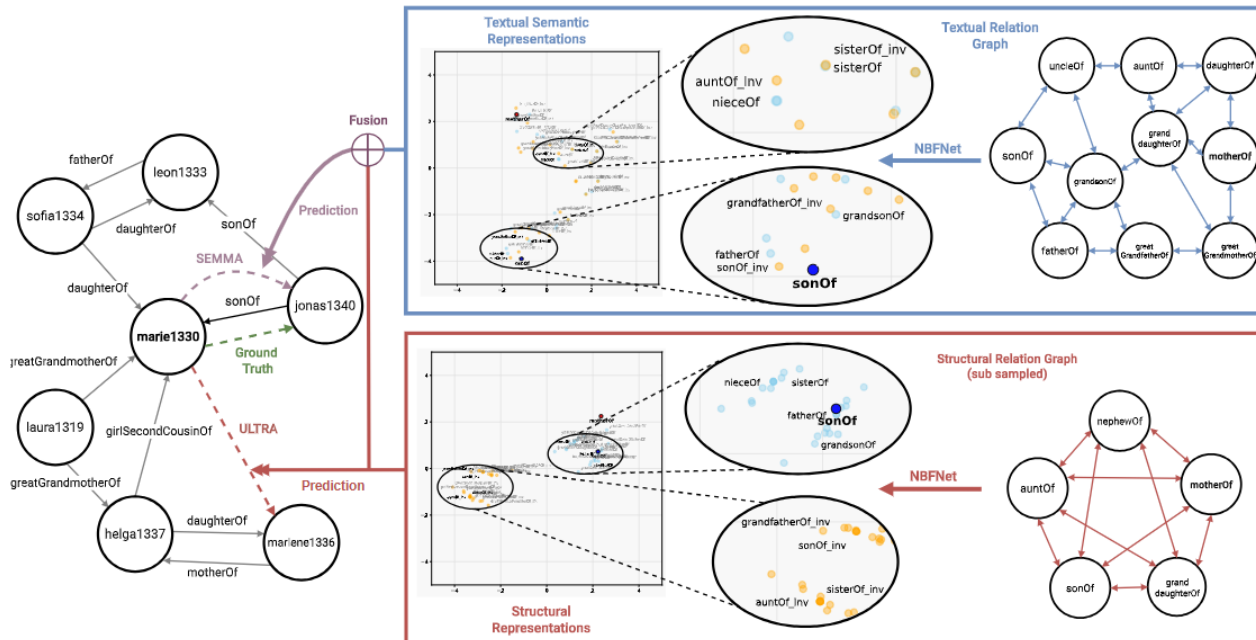
**Arvin^{dh} Arun^o ¹, Sumit Kumar^o ², Mojtaba Nayyeri ¹, Bo Xiong ³
Ponnurangam Kumaraguru ², Antonio Vergari ⁴, Steffen Staab ¹⁵**

¹Institute for AI, University of Stuttgart, ²IIT Hyderabad,
³Stanford University, ⁴University of Edinburgh, ⁵University of Southampton
arvin^{dh}.arun@ki.uni-stuttgart.de

LLM Integration Strategies

[2025][EMNLP][SEMMA]

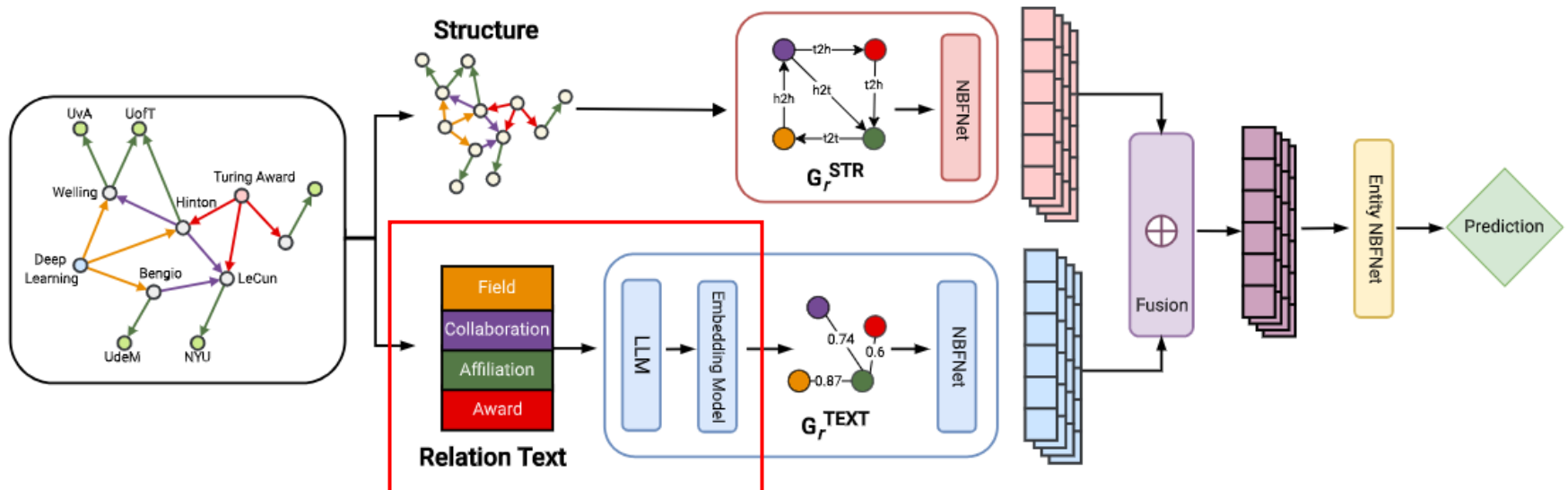
- 구조 신호만으로는 구분 자체가 불가능한 실패 사례
 - ✓ Query relation의 vocabulary가 test graph와 완전히 분리된 harder setting
 - ✓ 의미가 정반대인 두 relation을 구조만으로는 절대 구분할 수 없음
- 거의 모든 실세계 KG는 Text-Attributed KG(TAKG) — identifier에 자연어 라벨이 붙어 있음



LLM Integration Strategies

[2025][EMNLP][SEMMA]

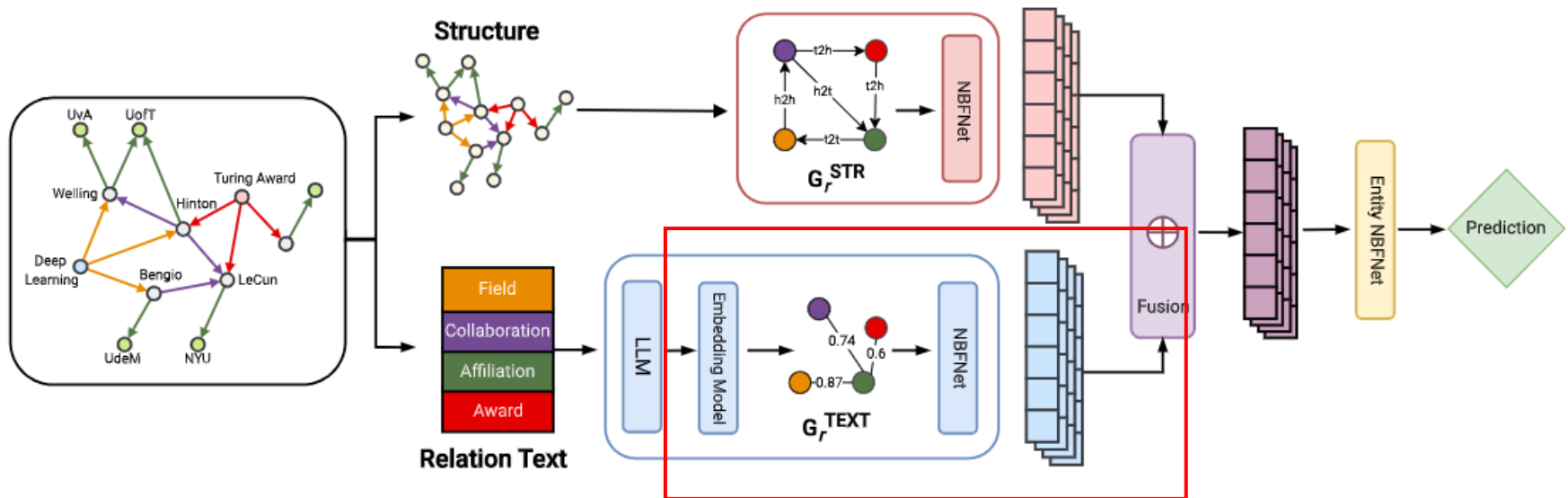
- **ULTRA와 동일한 구조에서, Relation Text Graph를 확장함**
- ① Extracting semantics from LLMs
 - ✓ 원본 relation label $\mathcal{T}_R$ 을 LLM에 넣어 정제된 이름 / Description 생성
 - ✓ 각 relation r 의 텍스트 임베딩 τ_r 을 계산



LLM Integration Strategies

[2025][EMNLP][SEMMA]

- **ULTRA와 동일한 구조에서, Relation Text Graph를 확장함**
- ② Textual Relation Graph 구축
 - ✓ 노드는 relation, edge는 텍스트 임베딩 간 cosine similarity를 가중치로 갖는 weighted edge
 - ✓ 모든 쌍을 연결하면 dense해져 over-smoothing이 발생하므로 필터링
 - ✓ 구조와 텍스트를 병렬로 처리 후 융합



CONTENTS

1. Background
2. Entity-Inductive Reasoning
3. Structural Signal-based KGFM
4. LLM Integration Strategies
5. Research Direction



Research Direction

Observation & Assumption

- LLM Based KGFM의 성능 관찰
 - ✓ 일부 데이터셋에서, LLM Based KGFM의 성능은 기초 모델인 ULTRA 보다 낮음
 - EX) MetaFarm (MRR+.0.2) / YAGO (MRR-.0.07)
- 데이터셋 별로 / 데이터셋 내에서 필요로 하는 Semantic 정보가 다름
 - ✓ 데이터셋 별 : 위에서 이미 관측된 사항 (SEMMA / KRLM)
 - ✓ 같은 데이터셋 내에서도 쿼리별로 필요로 하는 Semantic 정보가 다를 것이다
 - ✓ 구조적 정보가 충분한 경우에는 오히려 텍스트 정보가 노이즈가 되지 않을까
 - EX) fatherOf vs sonOf : 구조적으로 충분, 반대의 의미가 Semantic embedding 유사
 - ✓ 반대로 구조적 정보가 불충분한 경우거나, 구조적으로 구분이 불가능한 경우에는 텍스트 정보가 더 필요할 것이다

Research Direction

MOTIVATION : Background

- Structure-only KGFM은 Semantic Blindness를 가진다
 - ✓ 구조적으로 유사하지만 의미적으로 다른 relation (동일 구조 / 여러 의미)
 - ✓ Query Relation의 Support triple 부족 (Hard Inference)
- LLM-based KGFM의 등장 -> Semantic Blindness를 보완 -> 성능 향상 증명
 - ✓ SEMMA, PROLINK, KRLM은 relation name, description 또는 LLM의 사전지식을 이용해 그래프 구조만으로 구분하기 어려운 relation을 보완한다.
- 애초에 Knowledge Graph Foundation Model 자체는, 원래 Structure만 보는 것으로 설계
 - ✓ KGFM은 Feature 정보를 의도적으로 삭제함으로써, Unseen KG에 대해 일반화 용이 설계
 - ✓ 여기에 Semantic 정보를 포함하는 것은 의미가 없는 단어거나 구조적으로 충분한 KG에 대해 Robustness가 떨어지게 한다.
- 현재 Semantic 정보들이 어떤 상황에서 Fully-Inductive KG Reasoning에 도움이 되는지, 충분한 연구가 일어나지 않았음
 - ✓ 데이터셋 종류뿐만 아니라 쿼리 레벨에서도 요구하는 Semantic 정보량이 다르다
 - ✓ Structure만으로 충분히 구분 가능한 쿼리는, 오히려 Semantic 정보가 Knowledge Distortion 일으킨다

Analysis : INDEX

2-1) Semantic Utility 정의

2-2) 같은 데이터셋 안에서도 쿼리별로 Semantic Utility가 정말 다른가?

Semantic Utility : Background

1) A Theory of Usable Information Under Computational Constraints, ICLR 2020

- 제안 : 관찰자를 우리가 실제로 쓸 수 있는 함수들의 집합 $\mathcal{V}$ 로 제한
- Usable information은 두 손실의 차이 ("존재하는 정보" -> "쓸 수 있는 정보")

$$H_{\mathcal{V}}(Y | X) = \inf_{f \in \mathcal{V}} \mathbb{E}_{X,Y} [-\log f[X](Y)] \quad I_{\mathcal{V}}(X \rightarrow Y) = H_{\mathcal{V}}(Y | \emptyset) - H_{\mathcal{V}}(Y | X)$$

2) Understanding Dataset Difficulty with $\mathcal{V}$ -Usable Information, ICML 2022

- 조작 : inf를 제거하고, 실제로 학습해서 얻은 모델로 고정, 평균을 벗어나고 Instance 단위로 본다
- 데이터셋 난이도를 model-relative하게, 그리고 instance 단위로 분해해서 진단

$$\text{PVI}(x \rightarrow y) = -\log g[\emptyset](y) + \log g'[x](y)$$

3) Measuring Usable Information Beyond a Baseline, EMNLP 2021

- "X에 정보가 있느냐"가 아니라 "baseline B에는 없고 X에만 있는 정보가 있느냐"
- B를 이미 쓸 수 있는 상태에서, X가 추가로 제공하는 usable information.

$$I_{\mathcal{V}}(X \rightarrow Y | B) = \underbrace{H_{\mathcal{V}}(Y | B, \emptyset)}_{\text{B만 볼 때의 최선 loss}} - \underbrace{H_{\mathcal{V}}(Y | B, X)}_{\text{B와 X를 볼 때의 최선 loss}}$$

Research Direction

Semantic Utility : Definition

$$I_V(X \rightarrow Y) = H_V(Y | \emptyset) - H_V(Y | X)$$

$$I_V(X \rightarrow Y | B) = \underbrace{H_V(Y | B, \emptyset)}_{\text{B만 볼 때의 최선 loss}} - \underbrace{H_V(Y | B, X)}_{\text{B와 X를 볼 때의 최선 loss}}$$

$$\text{PVI}(x \rightarrow y) = -\log g[\emptyset](y) + \log g'[x](y)$$

$$\text{PVI}(x \rightarrow y | b); =; -\log g[b, \emptyset](y); +; \log g'[b, x](y)$$



**Semantic
Utility**

$$U(q) = \log p_{S+T}(t^* | q) - \log p_S(t^* | q)$$

Experiment1

Semantic

$$U(q) = \log p_{S+T}(t^* | q) - \log p_S(t^* | q)$$

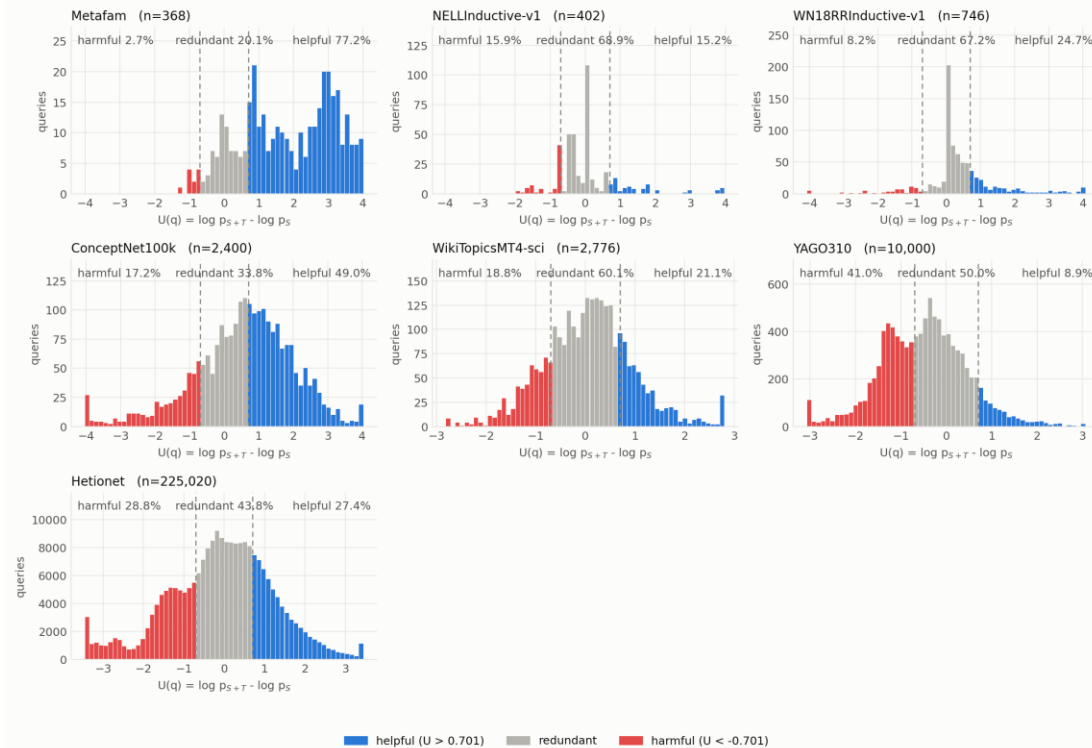
Query별 Semantic Utility 측정 실험

1. SEMMA를 Structure만으로 pretrain
2. SEMMA를 Structure + Text로 Pretrain
3. Inductive 상황에서 query별로 $U(q)$ 를 측정

Research Direction

Experiment1 결과

Fig A - query-level utility of the text channel, per KG



⇒ 핵심 관찰 : 같은 데이터셋 안에서도 쿼리에 따라서 Semantic Utility가 매우 다르며, 현재 LLM based KGFM은 이를 반영하지 못한다



Thank you for Listening