Fast Think-on-Graph: Wider, Deeper and Faster Reasoning of Large Language Model on Knowledge Graph

Xujian Liang [1,2] Zhaoquan Gu [2,3]

[1]Beijing University Of Posts And Telecommunications [2]Peng Cheng Laboratory [3]Harbin Institute of Technology (Shenzhen)

The 39th Annual AAAI Conference on Artificial Intelligence

Code is avaiable

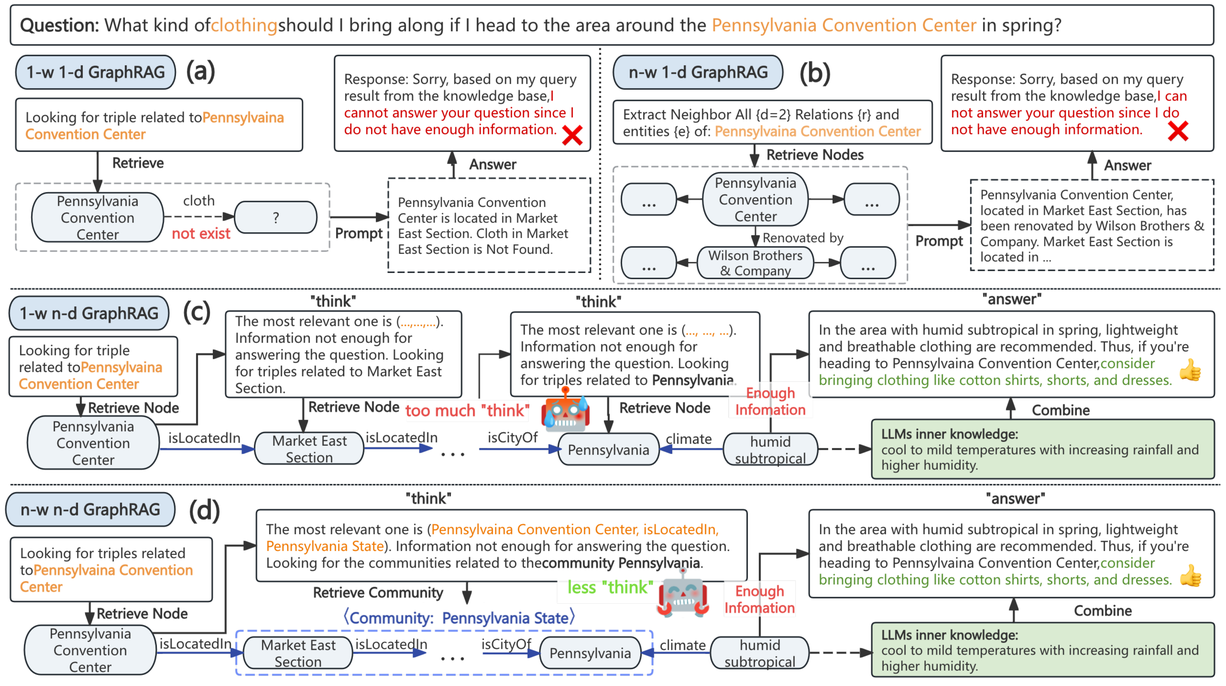

Figure 1. Comparison of 1-w 1-d Graph RAG (a), n-w Graph RAG (b), n-d Graph RAG (c) and n-w n-d GraphRAG (d)

Introduction

Retrieval-Augmented Generation (RAG) enhances LLMs by integrating external knowledge, but earlier methods face limitations. Naive RAG, relying on vector similarity, struggles with low precision, recall, and explainability due to embedding ambiguities. Graph-based RAG (GRAG) improves reasoning using knowledge graphs but suffers from computational inefficiency in n-d methods and reduced recall in dense graphs for n-w approaches (2). To address these, Fast Think-on-Graph (FastToG) introduces a novel n-d n-w paradigm, reasoning **"community by community"** via local community detection and pruning. FastToG improves accuracy, speeds up reasoning, and enhances explainability by converting graph structures into text for LLMs, outperforming prior methods. FastToG exhibited the following advantages:

1. **Higher Accuracy**: significant enhancement on the accuracy compared with the previous methods.
2. **Faster Reasoning**: notably shorten the reasoning chains and reduce the number of calls to the LLMs.
3. **Better Explainability**: The case study indicates that FastToG not only simplifies the retrieval for LLMs but also enhances the explainability for users.

Pipeline

1. Extract subject entities of the query (x) as a single-node start community (c_0).
2. Perform Community Detection on a subgraph to identify potential neighbor communities.
3. Apply coarse and fine pruning to refine and select the candidate communities (c_0^i) for reasoning chains.
   - **Modularity-based Coarse-Pruning**: based on [1], the modularity of a community c can be:

$$Q(c) = \sum \text{in} - \frac{(\sum \text{tot})^2}{2m} \tag{1}$$

   where the top k high-modularity communities C' can be:

$$C' := \text{argtopk}_{c \subseteq C} Q(c) \tag{2}$$

   - **LLMs-based Fine-Pruning**: FastToG prompts the LLMs to rank C':

$$C'' = \text{fine\_pruning}(x, C', \Pi, k) \tag{3}$$

4. Convert reasoning chains into text using methods like Graph2Text (G2T) or Triple2Text (T2T) for input to LLMs, iterating until an answer is generated or the process degrades to fallback methods.

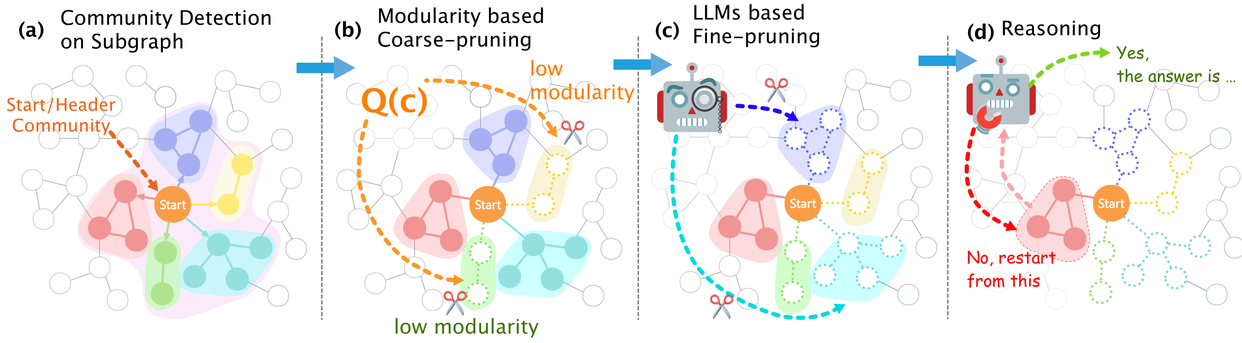

Figure 2. A general pipeline of the FastToG paradigm.

Experiment

- **Performance on Accuracy**: FastToG, which includes t2t and g2t mode, outperforms all previous methods. In particular, Ours(g2t) surpasses n-d 1-w (ToG) by 4.4% in Tab. 1.

Method	CWQ	WQSP	QALD	ZSRE	TREx	Creak
I/O Prompts	31.2	49.6	38.6	26.4	46.4	90.2
CoT	35.1	60.8	51.8	35.6	52.0	94.6
CoT-SC	36.3	61.2	52.4	35.8	52.0	95.0
1-d 1-w	35.5	59.2	50.7	39.4	56.1	92.0
1-d n-w	42.3	64.4	54.8	46.1	58.8	92.8
n-d 1-w	42.9	63.6	54.9	54.0	64.2	95.4
FastToG(t2t)	43.8	65.2	**56.1**	**54.4**	67.3	95.6
FastToG(g2t)	**45.0**	**65.8**	55.9	54.2	**68.6**	**96.0**

Table 1. Accuracy (%) for different datasets by gpt-4o-mini.

- **Performance on Efficiency**: With the growth of community size (fig 3), community-based ($MaxSize > 1$) reasoning can notably shorten the reasoning chains, reducing the number of calls to the LLMs.

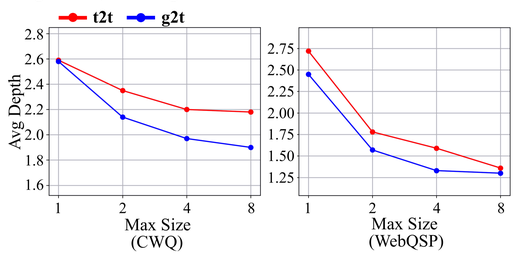

Figure 3. Average Depth versus Max size of community

References

[1] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008, 2008.

Case Study

The retrieval process of FastToG for the query "*Of the 7 countries in Central America, which consider Spanish an official language?*" highlights the efficiency of using communities over nodes for pruning. From fig 4, identifying *El Salvador* as an answer is easier through community connections, enhancing both LLM performance and user understanding.

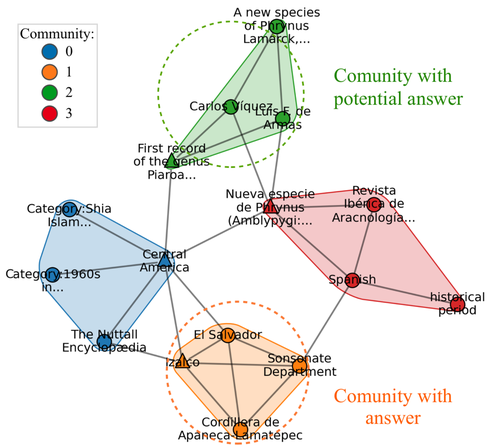

Figure 4. Visualization of the retrieval by FastToG

Think on Community	Think on Node
A. The Sonsonate Department is located in El Salvador, which is part of Central America. Spanish is the language used in the Sonsonate Department ...	A. El Salvador is belong to Central America
B. The Nuttall Encyclopida describes Mexico City as a city in Central America, which is part of North America. The Centralist Republic of Mexico ...	B. Nueva especie de Phrynus is main subject in Central America.
C. The first record of genus Piaroa Villarreal, Giupponi & Tourinho, 2008, is from Central America. Carlos Vquez, a Panamanian author, has written about ...	... more than 20 options Z. The South and Central American Club is held in Central America.

Table 2. Think on Community versus Think on Node