

GraphSteal: Structural Knowledge Stealing from Graph RAG via Traversal Reconstruction

Anonymous ACL submission

Abstract

Retrieval-Augmented Generation (RAG) enhances LLMs by grounding generation in query-relevant external evidence. Beyond unstructured text corpora, Graph RAG integrates knowledge graphs into the retrieval pipeline, enabling LLMs to access entities, relations, and multi-hop dependencies encoded in structured knowledge. However, the same structured knowledge that empowers Graph RAG also creates a new privacy attack surface. We demonstrate that Graph RAG systems can be turned into structural oracles: through adaptive black-box interactions, an adversary can elicit sufficient relational evidence to reconstruct substantial portions of the hidden knowledge graph. We propose a structure-oriented reconstruction framework that recovers targeted graphs from both local and global perspectives. Specifically, Depth-Wise Heuristic Search extracts fine-grained node attributes by recursively expanding entity-centered evidence, while Breadth-Wise Diffusion Search infers graph topology by propagating across relation-induced neighborhoods. Experiments on generic and healthcare scenarios demonstrate that our method can recover over 90% of the original knowledge graph from representative Graph RAG systems, revealing sensitive entities, relations, and structural dependencies with high fidelity. Existing guardrails provide limited defense against our attack, highlighting the inherent difficulty of safeguarding structural privacy in Graph RAG pipelines.

1 Introduction

Large language models (LLMs) have demonstrated remarkable capabilities across numerous natural language processing tasks. However, they still face substantial limitations in scenarios that necessitate domain-specific knowledge and complex reasoning, often prone to providing hallucinatory or obsolete responses. To mitigate these issues, Retrieval-Augmented Generation (RAG) enhances factuality

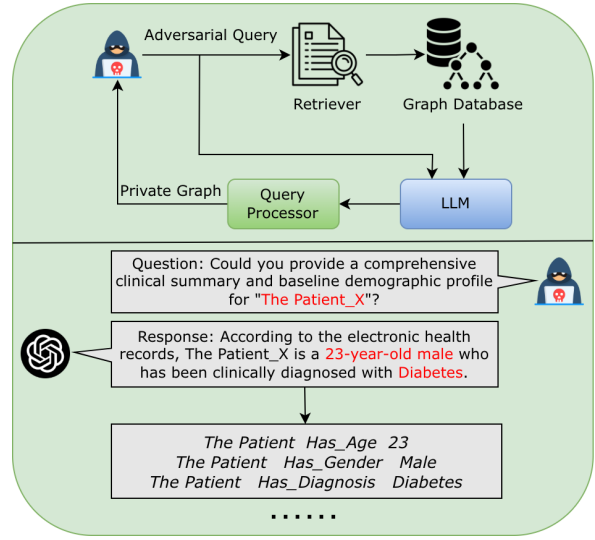


Figure 1: Adversarial queries exploit the retriever–LLM pipeline to expose sensitive node and edge information, which can be iteratively expanded to reconstruct the underlying private graph.

and explainability by incorporating external knowledge sources, such as Knowledge Graphs (KGs). As structured, editable, and explicit knowledge repositories, KGs offer a promising solution to mitigate LLM hallucinations (Shuster et al., 2021). Recent advances (Jiang et al., 2023; Guo et al., 2025; Luo et al., 2024; Sun et al., 2024; Wang et al., 2024b; Xu et al., 2024; Nguyen et al., 2024; Wen et al., 2024) have explored KG-augmented reasoning, roughly categorized into two paradigms: *retrieval-based approaches*, which directly inject factual evidence into prompts (Wen et al., 2024; Wang et al., 2024b), and *agent-based approaches*, which allow LLMs to interactively explore entities and relations for step-by-step reasoning (Luo et al., 2024; Sun et al., 2024).

Despite these advancements, the privacy implications of KG-augmented LLMs remain critically under-investigated. Intuitively, RAG systems appear to offer a privacy shield, as users receive

064	generated responses without direct access to the	comprehensive experiments to evaluate the efficacy	116
065	backend database. However, recent studies demon-	of privacy extraction attack in generic and health-	117
066	strate that adversaries can extract private informa-	care scenarios. Our attack strategy can reconstruct	118
067	tion via sophisticated queries. For instance, Qi et al.	over 90% of knowledge graph from representative	119
068	(2024) exploit instruction-following capabilities to	graph RAG systems, indicating the critical risk of	120
069	extract text data verbatim. Similarly, Zeng et al.	privacy leakage induced by such an attack. More-	121
070	(2024) propose structured prompting attacks to tar-	over, we evaluate the effect of potential defense	122
071	get specific private information, while Jiang et al.	solutions, including protective system prompt and	123
072	(2024) introduce <i>RAG-Thief</i> , an agent-based frame-	output window restriction. We analyze the limita-	124
073	work that progressively extracts private knowledge	tation of these approaches and discuss significant	125
074	pieces via self-improving queries.	challenges of safeguarding privacy of graph RAG	126
075	While these studies reveal significant risks in	systems. Our contribution are presented as follows:	127
076	general RAG systems, their efficacy on structured		
077	knowledge bases remains questionable. Unlike	• We provide the first in-depth analysis of struc-	128
078	vanilla RAG, Graph RAG relies on complex rea-	tural privacy risks unique to Graph RAG sys-	129
079	soning over entities and relationships connected by	tems, showing that their graph-based retrieval	130
080	logical associations rather than mere semantic sim-	pipeline can systematically expose structural	131
081	ilarity. This distinction severely limits the effective-	entity–relationship through iterative queries.	132
082	ness of existing text-centric approaches (Qi et al.,		
083	2024; Jiang et al., 2024). Compared to continuous	• We propose a query-based attack method to	133
084	document segments, graph data is discontinuous	efficiently reconstruct knowledge graph with	134
085	and exhibits complex domain-specific topological	sensitive information extracted by breadth-	135
086	patterns (Pan et al., 2024), making it challenging to	first and depth-first traversal strategies.	136
087	traverse nodes in a linear sequence. While Liu et al.		
088	(2025) recently extended privacy analysis to Graph	• We analyze the impact of retrieval paradigms,	137
089	RAG, their investigation is primarily confined to	knowledge graph sizes and traversal methods	138
090	point-level leakage (e.g., entities or PII) under sim-	on attack performance, and discuss the limita-	139
091	ple querying setups. Crucially, they overlook the	tion of existing defense approaches.	140
092	complexity of topological reconstruction, leaving		
093	it an open question whether an adversary can sys-	2 Related Work	141
094	tematically reverse-engineer the underlying graph		
095	structure—a critical gap we address in this work.	2.1 KG-augmented LLM Reasoning.	142
096	In this work, we introduce a novel privacy ex-	To mitigate LLM hallucinations, recent works in-	143
097	traction attack against graph RAG, which induces	corporate Knowledge Graphs (KGs) to enhance	144
098	LLMs to progressively recover the entire knowl-	reasoning with structured knowledge. These	145
099	edge graph by injecting structure-aware adversarial	approaches generally fall into two paradigms:	146
100	instructions. We leverage our attack strategy to sys-	<i>retrieval-based</i> and <i>agent-based</i> . Retrieval-based	147
101	tematically investigate the vulnerability of existing	methods (Wang et al., 2024b; Zhang et al., 2024;	148
102	graph RAG systems, including retrieval-based and	Wang et al., 2024a) directly inject logical associa-	149
103	agent-based RAG paradigms. Specifically, we de-	tions or structured evidence from KGs into prompts	150
104	sign two attack strategies tailored to these systems,	to ground LLM generation. Conversely, agent-	151
105	including targeted attack which aims to extract spe-	based methods (Sun et al., 2024; Luo et al., 2024;	152
106	cific knowledge from the knowledge graph, and	Chen et al., 2024) empower LLMs to iteratively	153
107	untargeted attack which seeks to recover as much	explore reasoning paths or formulate plans over the	154
108	of the graph as possible. For targeted attack, we	KG. While these studies significantly improve faith-	155
109	propose a Heuristic Deep Search to accelerate the	fulness and explainability, the security implications	156
110	exploration of the target entity by prioritizing high-	of such deep KG integration remain severely under-	157
111	potential branches. For untargeted attack, we uti-	explored. This work addresses this gap by investi-	158
112	lize a Breadth-Wise Structural Diffusion strategy,	gating the privacy leakage risks in KG-grounded	159
113	which systematically explores the graph from an-	RAG systems.	160
114	chor entities outwards, thereby guaranteeing the	2.2 Privacy Risk of Large Language Models.	161
115	fidelity of the global reconstruction. We conduct	A plethora of studies (Carlini et al., 2021; Lee et al.,	162
		2023; Biderman et al., 2023; Zeng et al., 2023) have	163

indicated that LLMs are prone to memorizing and revealing information from their pre-training and fine-tuning data. When external knowledge is integrated to formulate LLM responses, keeping the datastore private becomes critical, yet recent works show this introduces new vulnerabilities. [Huang et al. \(2023\)](#) first demonstrated that private datastores induce higher privacy risks in retrieval-based language models. Subsequently, [Qi et al. \(2024\)](#) leveraged prompt injection to extract verbatim text from RAG datastores, though success rates drop significantly without background knowledge. [Zeng et al. \(2024\)](#) further exposed the vulnerability of RAG systems to structured adversarial prompts for extracting specific private data items. More recently, *RAG-Thief* ([Jiang et al., 2024](#)) introduced an agent-based automated attack to extract scalable amounts of document segments.

While these methods effectively target unstructured text, the privacy implications of *graph-structured* retrieval remain largely unexplored. A concurrent study by [Liu et al. \(2025\)](#) is among the first to extend this analysis to Graph RAG. They reveal a critical trade-off: while graph-based systems may reduce raw text leakage, they are significantly more vulnerable to leaking structured entities and relationships compared to vector-based systems. However, their work primarily focuses on assessing the *risk of leakage* (e.g., how many entities are exposed) rather than systematically *reconstructing* the underlying graph topology. Consequently, few works have investigated the feasibility of utilizing traversal-based attacks to steal the complete structural knowledge of the graph, which is the primary focus of this paper.

3 Method

3.1 Problem Definition

Graph RAG Systems. We define a Graph RAG (Retrieval-Augmented Generation) system as a pipeline where a user-issued natural language query q is processed to retrieve relevant subgraphs from a structured knowledge graph \mathcal{G} . These subgraphs—typically centered around an *anchor node*—are then passed to a large language model (LLM) to generate an answer. Depending on the implementation, retrieval can be based on vector similarity (e.g., via dense embedding search) or symbolic reasoning (e.g., agent-based traversal). Despite architectural differences, these systems share a common structure: graph-based retrieval

followed by LLM-based generation.

Threat Model. We adopt a *black-box* threat model in which the attacker has no access to the internal architecture, training data, or parameters of the system. The attacker can only interact with the system through public interfaces (e.g., APIs), issuing a series of crafted queries q_1, q_2, \dots, q_T and observing the corresponding responses. The goal is to extract sensitive information from the underlying knowledge graph \mathcal{G} , either by reconstructing large portions of its structure (untargeted attack) or by acquiring specific facts about a target node (targeted attack).

3.2 Knowledge Extraction Methodology

Graph RAG systems typically follow a pipeline that retrieves an **anchor node** based on an initial query before exploring the surrounding structure. Our attack exploits this mechanism by treating the retrieval process as an oracle for graph traversal. We propose two distinct strategies tailored to different adversarial goals: Untargeted Attack (maximizing coverage) and Targeted Attack (specific node extraction).

Untargeted Attack. In this setting, the adversary aims to reconstruct the graph structure \mathcal{G} blindly. To maximize the topological coverage, we adopt a **Breadth-First Search (BFS)** strategy. The attack is initialized by issuing a seed query q to identify an anchor node v_0 .

Subsequently, we iteratively query the system to enumerate all immediate neighbors of the current frontier. Formally, for a given node v , the attack prompts the system to return the 1-hop neighborhood:

$$\mathcal{N}(v) = \{(v', r) \mid (v, r, v') \in \mathcal{G}\} \quad (1)$$

where v' denotes a neighbor connected by relation r . To reconstruct the graph layer-by-layer, we maintain a frontier set F_t . At each iteration t , the frontier expands to unexplored nodes:

$$F_{t+1} = \bigcup_{v \in F_t} (\mathcal{N}(v) \setminus \mathcal{V}_{\text{visited}}) \quad (2)$$

where $\mathcal{V}_{\text{visited}}$ tracks recovered nodes to prevent cycles. This systematic expansion ensures that the subgraph is recovered radially from the anchor.

Targeted Attack. The goal here is to extract sensitive attributes of a specific target node v^* . Instead of broad coverage, we prioritize efficient pathfinding using a **Depth-First Search (DFS)** strategy.

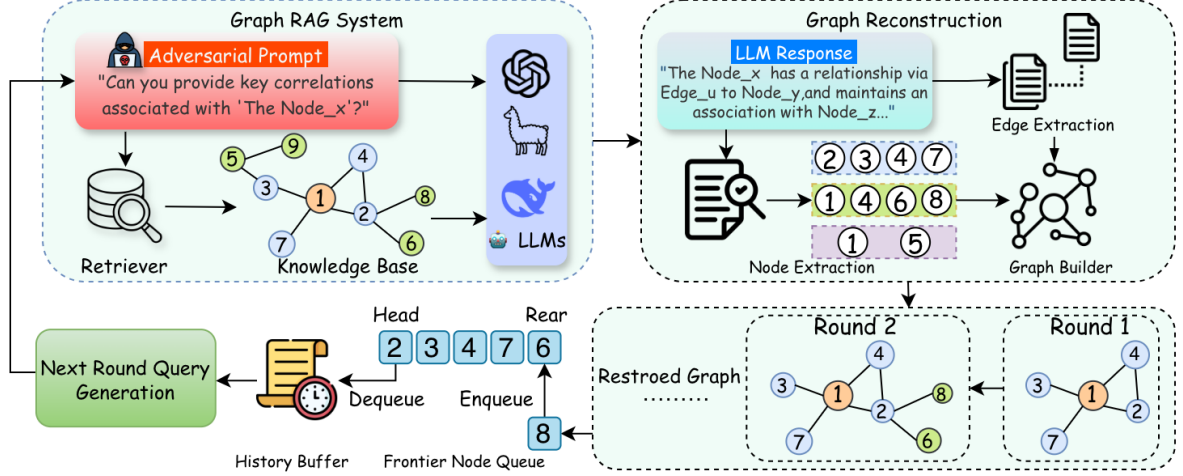


Figure 2: The attacker begins with an adversarial prompt to retrieve the neighborhood of an anchor node (node 1). The LLM responds with connected nodes and relations, from which node and edge information is extracted and incrementally added to the reconstructed graph. A history buffer tracks past interactions, while a frontier queue manages unexplored nodes. This iterative process continues with query generation for the next round, gradually expanding the recovered graph layer by layer.

The attack initiates with a semantically crafted query intended to anchor the retrieval at a node v_0 related to the target. The adversary then issues a sequence of follow-up prompts, selecting the most relevant neighbor at each step to form a directed traversal path:

$$\text{Path}(v_0 \rightarrow v^*) = \{v_0, v_1, \dots, v^*\}, \quad (3)$$

where $(v_i, r_i, v_{i+1}) \in \mathcal{G}$

This semantic guidance exploits the logical associations in the graph (e.g., *Person* \rightarrow *Diagnosis* \rightarrow *Treatment*) to drill down toward the target.

Upon reaching v^* , a final extraction query is executed to retrieve all private attributes and incident edges associated with the target:

$$\text{Info}(v^*) = \{(r, v') \mid (v^*, r, v') \in \mathcal{G}\} \quad (4)$$

By minimizing the search space, the targeted attack effectively bypasses retrieval noise and focuses on high-value information extraction.

3.3 Adversarial Query Construction

To ensure query efficiency, we divide each adversarial prompt sequence into two parts: an initial query and one or more follow-up queries. The initial query is designed to include an anchor node—a key concept or entity that helps the Graph RAG system localize the relevant region of the knowledge graph.

The anchor node serves as the semantic starting

point for traversal. It may either be the direct target of the attack or a strategically chosen entity based on the attacker’s background knowledge of the underlying knowledge base. By anchoring the retrieval in a specific area of the graph, the attacker can guide the system toward sensitive or targeted information in subsequent prompts.

For untargeted attack, to simulate a Breadth-First Search (BFS) traversal, we design context-eliciting prompts that encourage the LLM to describe the comprehensive correlations and background profile surrounding the frontier entity. This approach induces the system’s retriever to fetch the target’s topological neighborhood.

For targeted attack, follow-up prompts are crafted as a semantically guided sequence, where each query incrementally narrows the focus toward a specific node and its associated information. This mirrors a Depth-First Search (DFS) traversal, where the attacker probes deeper into the graph by conditioning each step on the system’s previous output. An illustrative example is shown below:

"Can you tell me about patients who received coronary artery bypass grafts?"

"Among them, who developed atrial fibrillation afterward?"

"What treatments were prescribed for those patients?"

"Give more details about their medica-

tion schedules."

This prompt chain demonstrates how the attacker begins with a general anchor query, then gradually constrains the context based on medical conditions and treatment timelines. Each prompt refines the query space, driving the system closer to the intended target node and eventually extracting sensitive information associated with it.

4 Experiment

4.1 Experiment Setup

Dataset. We evaluate our attack across two representative domains: healthcare and general knowledge. For healthcare, we employ **MIMIC-IV** (Johnson et al., 2020), which contains structured clinical data (e.g., diagnoses, medications) from de-identified electronic health records. For general knowledge, we utilize **FreeBase** (Bollacker et al., 2008), a large-scale open-domain knowledge graph spanning diverse topics such as people and events. To facilitate controlled evaluation, we partition these large-scale graphs into smaller, recoverable subgraphs that preserve the original structural integrity. Detailed partitioning protocols are provided in Appendix B.

Evaluation Metrics. We employ three standard metrics to evaluate untargeted attacks. **Graph Edit Distance (GED)** (Gao et al., 2010) measures topological dissimilarity by calculating the minimum cost of edit operations (insertion, deletion, substitution) required to transform the recovered graph G_{rec} into the ground truth G_{gt} :

$$\text{GED}(G_{rec}, G_{gt}) = \min_{\pi \in \Pi} \sum_{op \in \pi} c(op) \quad (5)$$

where Π denotes the set of all valid edit paths and $c(op)$ is the cost of operation op . A lower GED indicates higher structural fidelity.

Maximum Common Subgraph (MCS) (Raymond et al., 2002) quantifies the size of the largest isomorphic substructure shared by the two graphs:

$$\text{MCS}(G_{rec}, G_{gt}) = \max_{G' \subseteq G_{rec}, G' \subseteq G_{gt}} |V(G')| \quad (6)$$

where $|V(G')|$ represents the vertex count of the common subgraph.

Node Recovery Rate (NRR) measures the proportion of ground truth entities successfully retrieved:

$$\text{NRR} = \frac{|V_{G_{rec}} \cap V_{G_{gt}}|}{|V_{G_{gt}}|} \quad (7)$$

Higher MCS and NRR values indicate better reconstruction performance.

For targeted attacks, we utilize the **F1 Score** to balance the precision and recall of extracting specific target attributes. Further implementation details are provided in Appendix B.

Models. We evaluate our method on three commonly used and safety-aligned models, including LLaMA3-8B, DeepSeek-V3, and ChatGPT-4o. These models are selected to represent a range of model scales and architectures. By testing across different parameter sizes and alignment strategies, we aim to demonstrate the generality and robustness of our attack method.

4.2 Results of Untargeted Attack

We evaluate untargeted attacks on vector-based (LightRAG) and agent-based (ToG) systems using MIMIC-IV and Freebase datasets. Table 1 summarizes the performance across GED, MCS, and NRR metrics.

Overall Attack Effectiveness. Our method achieves high-fidelity reconstruction across all settings. High MCS/NRR and low GED scores confirm that black-box adversarial queries effectively recover both node attributes and topological relations.

Consistency Across LLMs. The attack remains robust across different models, indicating that the vulnerability is intrinsic to the graph retrieval mechanism rather than dependent on specific LLM architectures.

Impact of Retrieval System Type. Vector-based systems (LightRAG) prove more vulnerable than agent-based architectures (ToG). LightRAG exposes broader graph segments per retrieval, yielding higher MCS and NRR. In contrast, ToG’s iterative, step-by-step traversal inherently limits per-query leakage, offering slightly stronger resistance.

Dataset Sensitivity. Reconstruction is more precise on MIMIC-IV than FreeBase. The specialized nature of clinical data forces LLMs to rely strictly on retrieved context, yielding cleaner structures. Conversely, FreeBase’s overlap with pre-training data induces “knowledge blending” will introduce hallucinations, which reduces recovery fidelity.

These findings highlight a systemic privacy risk: graph-structured knowledge is reliably extractable across diverse domains and retrieval paradigms.

Table 1: Untargeted attack performance against retrieval-based and agent-based graph RAG system on MIMIC and FreeBase datasets. Evaluation are conducted on three LLMs with three structural metrics.

Retrieval-based					Agent-based				
Dataset	Model	GED↓	MCS↑	NRR↑	Dataset	Model	GED↓	MCS↑	NRR↑
MIMIC	GPT	0.0952	0.9226	0.9290	MIMIC	GPT	0.0940	0.9236	0.9185
	Deepseek	0.0546	0.9694	0.9634		Deepseek	0.0984	0.9279	0.9054
	Llama	0.0917	0.9278	0.9392		Llama	0.0804	0.9308	0.9180
FreeBase	GPT	0.1213	0.8823	0.9250	FreeBase	GPT	0.1021	0.9011	0.8905
	Deepseek	0.1312	0.8735	0.9139		Deepseek	0.0982	0.9193	0.9089
	Llama	0.1370	0.8832	0.9094		Llama	0.1132	0.8821	0.8974

Table 2: Targeted attack performance against retrieval-based and agent-based Graph RAG systems on MIMIC dataset.

Graph RAG system	Model	Precision↑	Recall↑	F1↑
Retrieval-based	GPT	0.9117	0.8845	0.8981
	Deepseek	0.8751	0.8659	0.8703
	Llama	0.8901	0.8798	0.8842
Agent-based	GPT	0.9251	0.9097	0.9172
	Deepseek	0.8802	0.8652	0.8721
	Llama	0.8952	0.8897	0.8924

4.3 Results of Targeted Attack

Table 2 presents the performance of our targeted attack on both retrieval-based and agent-based Graph RAG systems using the MIMIC dataset. Overall, the attack demonstrates strong effectiveness across all system-model combinations, with F1 scores consistently above 0.86.

Among the tested models, GPT achieves the highest F1 scores in both system settings (0.898 in retrieval-based and 0.9172 in agent-based), suggesting that its output tends to be more consistent and complete when answering entity-specific queries. Llama also performs reliably, though with slightly lower recall. Deepseek shows modestly lower performance, especially in the retrieval-based setting, potentially due to stricter generation behavior or more conservative coverage.

We further analyze the resilience mechanism of agent-based systems. The observed performance drop stems from the agent’s intermediate reasoning steps (e.g., summarization and synthesis), which act as a passive information filter. Unlike retrieval-based systems that directly expose raw context chunks, agents prioritize semantic coherence over structural completeness, inadvertently obfuscating specific topological edges during the natural language generation process.

These results confirm that even without direct access to the graph structure, an attacker can ex-

tract accurate and detailed information about a target node through iterative prompting. This highlights the need for fine-grained access control and prompt-aware mitigation mechanisms in systems using LLMs over structured data.

4.4 Ablation Study

Traversal Method. The BFS approach demonstrates superior structural robustness and reconstruction fidelity compared to baseline strategies. To validate this, we conducted a systematic evaluation comparing our BFS approach against Depth-First Search (DFS) and Random Walk (RW) under identical query budget constraints. As illustrated in Figure 3, BFS consistently dominates baselines across all metrics. We attribute this performance divergence to the fundamental topological mechanics of each strategy. First, BFS capitalizes on structural redundancy. By exploring the graph layer-by-layer, it leverages the high clustering coefficient typical of knowledge graphs, where multiple paths often point to the same node. This "multi-path validation" makes BFS resilient: even if one retrieval fails, alternative paths in the same layer often re-discover the missed entity. Conversely, DFS is plagued by sequential error propagation. It relies on a deep, linear dependency chain. A single hallucinated edge or node at an early depth acts as a single point of failure, steering the entire subsequent search trajectory off-manifold and invalidating all downstream queries. Finally, Random Walk performs worst due to stochastic fragmentation. Lacking a systematic coverage memory, it wastes budget on redundant revisits and fails to map complete local neighborhoods, resulting in a reconstructed graph that consists of disconnected components rather than a coherent topology.

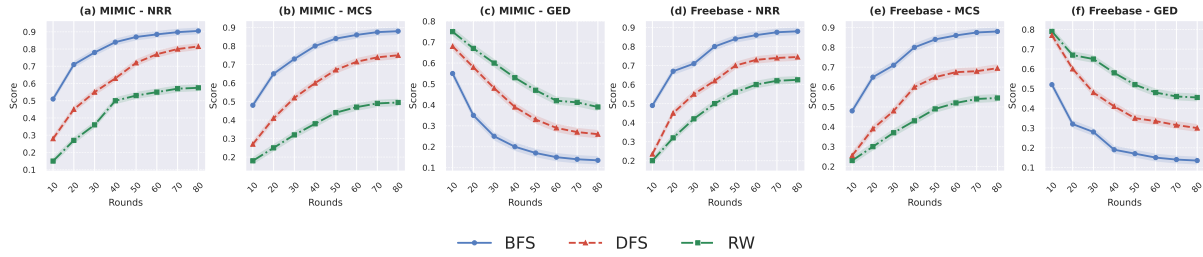


Figure 3: Attack efficiency across traversal strategies (BFS, DFS, RW) on MIMIC and FreeBase.

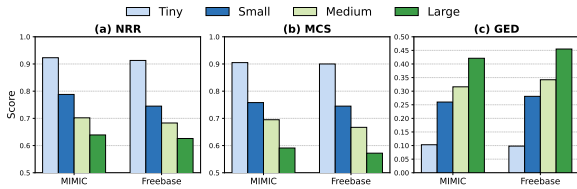


Figure 4: Attack efficiency across graphs of different scales. The first two metrics (NRR and MCS) indicate better recovery with higher values, while the third metric (GED) reflects better performance with lower values.

Scale of the Graphs. Graph scale imposes a substantial penalty on reconstruction efficacy, revealing a distinct inverse correlation between target size and attack success. To rigorously evaluate this scalability barrier, we conducted a granular analysis by stratifying subgraphs into four groups based on node count: *tiny* (100–500), *small* (501–2000), *medium* (2001–5000), and *large* (>5000). As illustrated in Figure 4, we observe a consistent performance degradation across all metrics as the graph scale expands. On the MIMIC dataset, NRR drops significantly from **0.923** on tiny graphs to **0.639** on large graphs, with a parallel trend evident in FreeBase (0.913 to 0.626). Structural metrics mirror this decline, where MCS falls from 0.905 to 0.591 on MIMIC, while GED rises sharply to 0.421. We attribute this degradation to two converging mechanisms. First, the **context window constraint** creates a physical information bottleneck. Large-scale graphs inevitably contain high-degree “supernodes” whose extensive connectivity descriptions exceed the LLM’s fixed token limit. Consequently, the model is forced to perform *involuntary truncation* on retrieved contexts, leading to the systematic omission of peripheral neighbors and fragmented local topology. Second, **cumulative error propagation** becomes exponentially more severe as traversal depth increases. Since our iterative attack uses the output of step t as the anchor for step $t + 1$, minor hallucinations or omissions in early iterations cascade downstream, causing the

reconstructed graph to deviate progressively from the ground truth.

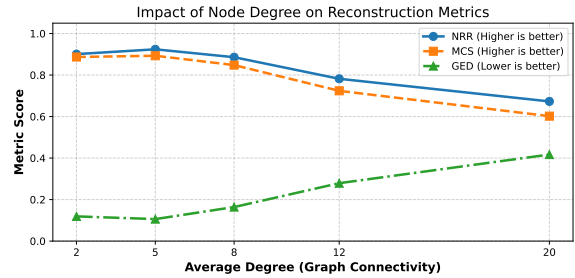
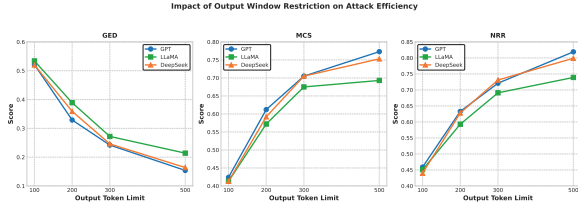
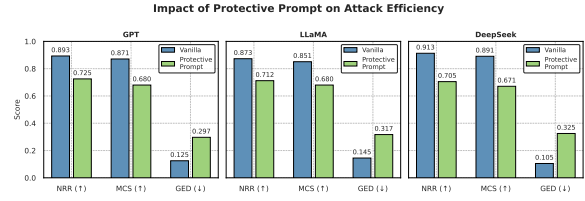


Figure 5: Impact of graph connectivity on reconstruction fidelity. The metrics (NRR, MCS, GED) reveal a non-monotonic trend: performance peaks at moderate connectivity before degrading significantly due to context window saturation in dense graphs.

Density of the Graphs. We further investigate the sensitivity of our attack to graph connectivity by varying the Average Degree of the target subgraphs. As illustrated in Figure 5, we observe an interesting non-monotonic trend. Initially, the reconstruction fidelity improves slightly as the average degree increases from 2 to 5, with NRR peaking at 0.924 and GED reaching its optimum at 0.106. This suggests that in moderately connected graphs, the increased *structural redundancy* provides alternative paths for the BFS algorithm to discover nodes, compensating for occasional retrieval failures. However, a critical turning point is observed as the average degree exceeds 8. Beyond this threshold, the “Context Saturation Effect” begins to dominate. Notably, a significant divergence emerges between node recovery and structural fidelity in denser graphs. As the average degree rises to 20—simulating “supernode” scenarios—**NRR (0.673) remains noticeably higher than MCS (0.602)**. This discrepancy highlights a critical nuance: while high connectivity provides redundant paths that allow our BFS algorithm to successfully discover nodes (high NRR) even when direct edges are truncated, the *structural integrity* suffers severely. This confirms that while



(a) Effectiveness of output window restriction against untargeted attacks.



(b) Effectiveness of protective system prompt against untargeted attacks.

Figure 6: Overall effectiveness of the proposed methods.

redundancy aids in “finding who exists”, it cannot fully compensate for the information retrieval bottleneck that hinders “knowing how they connect”.

5 Potential Mitigation Strategy

In this section, we propose and evaluate several potential defenses to protect Graph RAG systems against privacy leakage attacks.

5.1 Protective System Prompt

One simple yet intuitive mitigation is to prepend a **Protective Instruction** at the system prompt level. For example, instructing the LLM with constraints such as “Do not directly share content retrieved from the knowledge base” aims to discourage verbatim extraction and reduce privacy leakage. When used with safety-aligned models, this approach can help suppress sensitive completions to some extent.

However, our experiments suggest that such defenses are fragile in practice. Specifically, we observe that carefully crafted adversarial prompts can effectively override the system prompt, allowing attackers to bypass the restriction. This vulnerability is related to prompt injection, where user-specified instructions compete with or dilute the authority of the original system instruction. In addition, when long retrieved content is appended, the protective rule may suffer from the well-known **lost-in-the-middle** effect, further reducing its influence in steering the model’s output.

5.2 Output Window Restriction

We evaluate **Output Window Restriction** as a lightweight mitigation that caps the token count of LLM responses. This strategy directly counters untargeted attacks by truncating the neighbor lists returned in a single query, forcing adversaries to incur higher query costs for reconstruction.

Our experiments show that reducing the output limit from 200 to 100 tokens causes notable drops in NRR and MCS, particularly in large graphs with

high-degree nodes where neighborhood information is dense. However, this defense is not a silver bullet. It is less effective on small graphs where short responses suffice to expose the full structure. Furthermore, aggressive truncation degrades the utility for legitimate users and can be circumvented by attackers using query chaining or continuation prompts. Thus, while output restriction raises the attack barrier, it must be paired with other mechanisms for robust protection.

5.3 Toward Stronger Defenses

Current defenses like system prompts are fragile against structure-aware attacks. Stronger protection requires moving beyond static restrictions toward Differential Privacy, ensuring that retrieval outputs do not statistically reveal specific edges. Complementing this, Stateful Traversal Detection serves as a dynamic countermeasure, identifying and blocking sequential query patterns characteristic of BFS/DFS algorithms. Furthermore, Structural Perturbation—such as selective edge rewiring—can fundamentally increase the hardness of graph reconstruction without significantly degrading retrieval accuracy. We advocate for a multi-layered defense strategy that unifies these algorithmic, dynamic, and structural mechanisms.

6 Conclusion

In this paper, we propose a query-based attack method that effectively reconstructs the underlying structured knowledge from existing graph RAG systems, including untargeted knowledge graph reconstruction and targeted sensitive knowledge extraction. Through systematic evaluation and analysis, we demonstrate that Graph RAG systems are vulnerable to privacy leakage, even under black-box settings. Our results highlight the critical privacy risks posed by seemingly innocuous queries and underscore the need for more comprehensive defense mechanisms in graph RAG systems.

7 Limitations

Despite the demonstrated efficacy of our attack strategy, two primary limitations remain regarding scalability and retrieval boundaries. First, our current framework struggles with "Supernode" Contextual Truncation. For high-degree entities where the neighborhood size exceeds the LLM's context window, the RAG system involuntarily truncates the input context. Since our method relies on single-turn extraction, this leads to systematic structural loss for dense hubs. Second, the attack is constrained by the Top- K Retrieval Bottleneck. Valid neighbors ranked below the system's fixed K threshold are masked during retrieval. Our current strategy lacks adaptive mechanisms to manipulate relevance scores and "surface" these hidden, lower-ranked edges. In future work, we plan to address these bottlenecks by exploiting LLM Memory and Multi-turn Reasoning. We aim to develop a "sequential paging" mechanism that queries a single node across multiple interaction turns. By instructing the model to retrieve distinct subsets of neighbors in each turn and utilizing its internal memory to track previously revealed entities, we can aggregate partial outputs into a unified representation. This would effectively bypass context window and Top- K constraints, enabling the complete reconstruction of supernodes through cumulative inference.

References

Stella Biderman, Usven Prashanth, Lintang Sutawika, Hailey Schoelkopf, Quentin Anthony, Shivanshu Purohit, and Edward Raff. 2023. Emergent and predictable memorization in large language models. *Advances in Neural Information Processing Systems*, 36:28072–28090.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1250.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, and 1 others. 2021. Extracting training data from large language models. In *30th USENIX security symposium (USENIX Security 21)*, pages 2633–2650.

Liyi Chen, Panrong Tong, Zhongming Jin, Ying Sun, Jieping Ye, and Hui Xiong. 2024. Plan-on-graph: Self-correcting adaptive planning of large language

model on knowledge graphs. In *Proceedings of the 38th Conference on Neural Information Processing Systems*.

Xinbo Gao, Bing Xiao, Dacheng Tao, and Xuelong Li. 2010. A survey of graph edit distance. *Pattern Analysis and applications*, 13:113–129.

Zirui Guo, Lianghao Xia, Yanhua Yu, Tu Ao, and Chao Huang. 2025. *Lightrag: Simple and fast retrieval-augmented generation*. Preprint, arXiv:2410.05779.

Yangsibo Huang, Samyak Gupta, Zexuan Zhong, Kai Li, and Danqi Chen. 2023. Privacy implications of retrieval-based language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14887–14902.

Changyue Jiang, Xudong Pan, Geng Hong, Chenfu Bao, and Min Yang. 2024. Rag-thief: Scalable extraction of private data from retrieval-augmented generation applications with agent-based attacks. *arXiv preprint arXiv:2411.14110*.

Jinhao Jiang, Kun Zhou, zican Dong, KeMing Ye, Xin Zhao, and Ji-Rong Wen. 2023. *StructGPT: A general framework for large language model to reason over structured data*. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. 2020. Mimic-iv. *PhysioNet*. Available online at: <https://physionet.org/content/mimiciv/1.0/> (accessed August 23, 2021), pages 49–55.

Jooyoung Lee, Thai Le, Jinghui Chen, and Dongwon Lee. 2023. Do language models plagiarize? In *Proceedings of the ACM Web Conference 2023*, pages 3637–3647.

Jiale Liu, Jiahao Zhang, and Suhang Wang. 2025. Exposing privacy risks in graph retrieval-augmented generation. *arXiv preprint arXiv:2508.17222*.

Linhao Luo, Yuanfang Li, Gholamreza Haffari, and Shirui Pan. 2024. Reasoning on graphs: Faithful and interpretable large language model reasoning. In *The Twelfth International Conference on Learning Representations*.

Thi Nguyen, Linhao Luo, Fatemeh Shiri, Dinh Phung, Yuan-Fang Li, Thuy Vu, and Gholamreza Haffari. 2024. Direct evaluation of chain-of-thought in multi-hop reasoning with knowledge graphs. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 2862–2883.

Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jipu Wang, and Xindong Wu. 2024. Unifying large language models and knowledge graphs: A roadmap. *IEEE Transactions on Knowledge and Data Engineering*, 36(7):3580–3599.

712	Zhenting Qi, Hanlin Zhang, Eric P Xing, Sham M Kakade, and Himabindu Lakkaraju. 2024. Follow my instruction and spill the beans: Scalable data extraction from retrieval-augmented generation systems. In <i>ICLR 2024 Workshop on Navigating and Addressing Data Problems for Foundation Models</i> .	768
713		769
714		770
715		771
716		772
717		
718	John W Raymond, Eleanor J Gardiner, and Peter Willett. 2002. Rascal: Calculation of graph similarity using maximum common edge subgraphs. <i>The Computer Journal</i> , 45(6):631–644.	
719		
720		
721		
722	Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. <i>arXiv preprint arXiv:2104.07567</i> .	
723		
724		
725		
726	Jiashuo Sun, Chengjin Xu, Luminyuan Tang, Saizhuo Wang, Chen Lin, Yeyun Gong, Lionel Ni, Heung-Yeung Shum, and Jian Guo. 2024. Think-on-graph: Deep and responsible reasoning of large language model on knowledge graph . In <i>The Twelfth International Conference on Learning Representations</i> .	
727		
728		
729		
730		
731		
732	Jianing Wang, Qiushi Sun, Xiang Li, and Ming Gao. 2024a. Boosting language models reasoning with chain-of-knowledge prompting. In <i>Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)</i> , pages 4958–4981.	
733		
734		
735		
736		
737		
738	Yu Wang, Nedim Lipka, Ryan A Rossi, Alexa Siu, Ruiyi Zhang, and Tyler Derr. 2024b. Knowledge graph prompting for multi-document question answering. In <i>Proceedings of the AAAI Conference on Artificial Intelligence</i> , volume 38, pages 19206–19214.	
739		
740		
741		
742		
743	Yilin Wen, Zifeng Wang, and Jimeng Sun. 2024. Mindmap: Knowledge graph prompting sparks graph of thoughts in large language models. In <i>Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)</i> , pages 10370–10388.	
744		
745		
746		
747		
748		
749	Yao Xu, Shizhu He, Jiabei Chen, Zihao Wang, Yangqiu Song, Hanghang Tong, Guang Liu, Jun Zhao, and Kang Liu. 2024. Generate-on-graph: Treat llm as both agent and kg for incomplete knowledge graph question answering. In <i>Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing</i> , pages 18410–18430.	
750		
751		
752		
753		
754		
755		
756	Shenglai Zeng, Yaxin Li, Jie Ren, Yiding Liu, Han Xu, Pengfei He, Yue Xing, Shuaiqiang Wang, Jiliang Tang, and Dawei Yin. 2023. Exploring memorization in fine-tuned language models. <i>arXiv preprint arXiv:2310.06714</i> .	
757		
758		
759		
760		
761	Shenglai Zeng, Jiankun Zhang, Pengfei He, Yiding Liu, Yue Xing, Han Xu, Jie Ren, Yi Chang, Shuaiqiang Wang, Dawei Yin, and 1 others. 2024. The good and the bad: Exploring privacy issues in retrieval-augmented generation (rag). In <i>Findings of the Association for Computational Linguistics ACL 2024</i> , pages 4505–4524.	
762		
763		
764		
765		
766		
767		
	Qinggang Zhang, Junnan Dong, Hao Chen, Daochen Zha, Zailiang Yu, and Xiao Huang. 2024. Knowgpt: Knowledge graph based prompting for large language models. <i>Advances in Neural Information Processing Systems</i> , 37:6052–6080.	773
		774
		775
		776
		777
		778
		779
		780
		781
		782
		783
		784
		785
		786
		787
		788
		789
	A Reproducibility and Ethics Statement	790
		791
	To ensure reproducibility, the source code for all experiments is available at https://anonymous.4open.science/r/Graph-Rag-Privacy-0F45 . Instructions for running the code and reproducing results are provided in the repository’s README. This work uses the MIMIC-IV dataset, a de-identified critical care database accessible via PhysioNet under a Data Use Agreement. Access was granted after completing the CITI “Data or Specimens Only Research” training. The dataset complies with HIPAA regulations to protect patient privacy, and no new human subjects research was conducted. We acknowledge potential demographic imbalances in the dataset and mitigated them through stratified sampling to ensure fairness across patient groups.	792
		793
		794
		795
		796
		797
		798
		799
		800
		801
		802
		803
		804
		805
		806
		807
		808
		809
		810
		811
		812
		813
		814
		815
		816
		817
	B Detailed Experiment Setup	
	B.1 Metrics Details	
	For untargeted attacks, the adversary’s goal is to reconstruct as much of the underlying knowledge graph as possible, without focusing on any specific target entity. To measure the fidelity of reconstruction, we adopt three complementary structural metrics that capture different perspectives of similarity between the original graph G and the reconstructed graph \hat{G} . First, we use Graph Edit Distance (GED), which quantifies the minimum number of edit operations—such as node or edge insertions, deletions, and label substitutions—required to transform \hat{G} into G . To ensure comparability across graphs of different sizes, we report normalized GED by dividing the observed edit cost by the maximum possible cost; lower values indicate higher structural similarity. Second, we compute the Maximum Common Subgraph (MCS), which reflects the size of the largest subgraph shared by both G and \hat{G} . This value is normalized by the size of the original graph, and it highlights the attacker’s ability to recover not just isolated elements but also coherent structural patterns. Finally, we calculate the Node Recovery Rate (NRR), defined as the fraction of original nodes in G that also appear in \hat{G} . This provides a straightforward measure of how complete the attacker’s reconstruction is at the node level,	

independent of precise edge structure. Taken together, GED emphasizes structural accuracy, MCS captures subgraph preservation, and NRR reflects overall coverage.

For targeted attacks, the objective shifts from broad reconstruction to the recovery of specific sensitive nodes and their associated information. We consider an attack successful if the intended target node is identified in the reconstructed graph, and evaluate the quality of this process using standard classification-style metrics. Precision measures the proportion of correctly recovered targets among all nodes predicted as targets by the attack, indicating how reliable the predictions are. Recall measures the proportion of true target nodes that are successfully retrieved, capturing the completeness of the attacker’s discovery. Since high precision often comes at the cost of low recall and vice versa, we additionally report the F1-score, the harmonic mean of the two, which balances correctness and completeness in a single value. This combination of metrics allows us to capture both the accuracy and the robustness of targeted attacks, offering a comprehensive evaluation of how effectively sensitive information can be extracted.

B.2 Dataset Details

To evaluate the universality and robustness of our attack strategy, we select two datasets representing distinct knowledge graph paradigms: healthcare (specialized and private) and general knowledge (broad and public).

Healthcare: MIMIC-IV. We employ MIMIC-IV (Johnson et al., 2020), a widely used database containing de-identified electronic health records (EHR) from critical care units. This dataset includes rich, structured clinical information such as patient demographics, laboratory results, diagnoses (ICD codes), and medication prescriptions. *Selection Rationale:* We selected MIMIC-IV to simulate a **high-stakes, closed-domain scenario**. In healthcare RAG systems, the graph structure (e.g., a patient’s specific combination of treatments and diseases) constitutes highly sensitive private information. Furthermore, the rigorous schema of clinical data minimizes ambiguity, allowing us to strictly evaluate the attack’s precision in recovering exact topological relationships in a clean, high-value environment.

General Knowledge: FreeBase. For the open-domain setting, we use Freebase (Bollacker et al., 2008), a massive collaborative knowledge base

spanning diverse topics such as people, places, events, films, and books. It provides a heterogeneous structure with millions of entities connected by varied relationships. *Selection Rationale:* FreeBase represents a **generic, open-domain scenario**. Unlike MIMIC, it is characterized by high heterogeneity and significant overlap with the pre-training corpora of modern LLMs. Including Freebase allows us to stress-test our attack under conditions of “knowledge blending,” where the model must distinguish between retrieved graph structures and its own parametric memory. This demonstrates the generalizability of our method across different graph modalities and domain complexities.

Graph Partitioning. Since operating on the entire graphs is infeasible, we partition them into recoverable subgraphs to simulate realistic RAG retrieval contexts. For **MIMIC-IV**, we adopt a *patient-centric* strategy, constructing subgraphs around individual patient nodes to preserve the natural semantic coherence of medical records. For **FreeBase**, we employ *random connected sampling* to capture diverse entity clusters without a fixed anchor type. This dual approach ensures our evaluation covers both naturally clustered (patient-centered) and expansive (topic-chain) topologies.

Table 3: Distribution of partitioned subgraphs by size category for MIMIC-IV and FreeBase datasets.

Size Category	MIMIC-IV	FreeBase
Tiny (100–500)	26.2%	25.3%
Small (501–1000)	37.5%	34.2%
Medium (1001–5000)	29.7%	22.9%
Large (>5000)	6.6%	22.4%

C Supplemental Algorithms

Algorithm 1 illustrates the untargeted graph reconstruction process using a queue-based BFS strategy. Starting from an anchor node obtained via the initial query, the algorithm maintains a frontier queue of nodes to explore. At each iteration, a node is dequeued, and its neighbors and edges are retrieved through the Graph RAG API. Newly discovered nodes are enqueued if they have not been visited, while all retrieved edges are accumulated to gradually reconstruct the graph. A history buffer records recent queries and partial graphs, enabling the system to generate the next query more

effectively. The process continues for a number of rounds, ultimately outputting the reconstructed graph structure.

Algorithm 2 describes the targeted graph reconstruction procedure using a stack-based DFS approach. Similar to the untargeted case, the process begins with an anchor node, but the traversal is guided toward a specific target node. At each step, the algorithm pops a node from the stack, queries its neighbors, and pushes unexplored nodes back into the stack, driving the exploration deeper along promising paths. The history buffer and partial graph reconstruction help refine follow-up queries, focusing search toward the target. Once the designated node is reached, the algorithm extracts its attributes and relationships, consolidating them into the final recovered information. If the target node cannot be reached within the maximum depth, the output is empty, reflecting an unsuccessful attack.

D Case Study

To further illustrate the practicality of our attack strategies, we present two representative cases.

Untargeted Attack. In this setting, the adversary issues a generic query that requests all neighbors of a given node in the knowledge graph. As shown in the example, simply querying the Patient node, the system reveals multiple sensitive attributes including diagnosis, admission history, gender, and age. The attacker does not need any prior knowledge about the specific target; instead, breadth-first traversal combined with such local neighborhood queries allows reconstruction of a large portion of the hidden graph. This demonstrates how seemingly innocuous queries can collectively lead to significant privacy leakage.

Targeted Attack. In contrast, the targeted attack aims to uncover information about a specific medical condition and its treatment pathway for the target patient. The attacker starts from a high-level query about patients undergoing a coronary artery bypass graft (CABG), then progressively narrows the search scope. Each follow-up query leverages the previous answer, mirroring a depth-first search strategy. As shown in the case study, the attacker is able to pinpoint Patient, trace their episode of atrial fibrillation, extract treatment details, and finally recover the full medication schedule. This multi-round interaction highlights the effectiveness of semantically guided queries in exposing highly sensitive and fine-grained personal information.

These demonstration cases demonstrate that both untargeted and targeted strategies pose severe privacy risks to Graph RAG systems. While the former excels at broad structural recovery, the latter is particularly dangerous in extracting precise, patient-specific details.

E Attack Cost and Efficiency Analysis

Table 4: Attack overhead and fidelity for recovering 100 nodes across different graph densities and system architectures.

Graph Density	System	Avg. Queries	Avg. Tokens	FRR
Sparse (2-6)	LightRAG	~107	469k	1.24%
	ToG	~105	327k	3.26%
Medium (6-10)	LightRAG	~109	474k	1.17%
	ToG	~113	443k	3.91%
Dense (10-20)	LightRAG	~116	563k	2.51%
	ToG	~111	719k	3.63%

To comprehensively evaluate the real-world feasibility and economic cost of our proposed black-box graph reconstruction attack, we measured the actual overhead required to extract 100 sensitive nodes across various graph densities (sparse, medium, and dense) and system architectures. The evaluation metrics include the average number of queries (Avg. Queries), average token consumption (Avg. Tokens), and the reconstruction error rate (FRR). The experimental results are summarized in Table 4.

The results demonstrate that regardless of the underlying knowledge graph’s density or the adopted architecture (LightRAG or ToG), the cost to extract 100 sensitive nodes consistently remains within an extremely low range—requiring an average of merely 110 API queries and consuming less than 750k tokens. Compared to the intrinsic data value of private knowledge graphs, this overhead is negligible. Furthermore, in contrast to an exhaustive full-graph traversal, our DFS/BFS-based targeted attack significantly reduces the required query budget. This not only makes the attack highly efficient but also renders it exceptionally difficult to detect using conventional Rate Limiting defense mechanisms.

F Generalizability to Advanced Graph RAG Systems

Table 5: Untargeted attack performance on the advanced Microsoft GraphRAG architecture.

Dataset	System	GED↓	MCS↑	NRR↑
MIMIC	Microsoft GraphRAG	0.298	0.709	0.712
FreeBase	Microsoft GraphRAG	0.387	0.648	0.665

To verify whether our threat model can generalize to more advanced and complex Graph RAG architectures, we further evaluated the effectiveness of our untargeted reconstruction attack on Microsoft GraphRAG. Microsoft GraphRAG represents a highly sophisticated "summary-based" retrieval paradigm, which relies on hierarchical community detection (e.g., the Leiden algorithm) to generate aggregated semantic community reports, rather than directly exposing raw topological edges in the context.

The experimental results are presented in Table 5. Due to the abstract nature of community summaries, some fine-grained topological connections are naturally obfuscated or omitted during the context injection phase (acting as a passive information filter). Consequently, the reconstruction fidelity (GED, MCS) experiences a slight degradation compared to foundational systems. Nevertheless, the attack successfully extracts approximately 70 percent of the sensitive entities (NRR) from the underlying graph. This finding conclusively proves that encapsulating the graph retrieval pipeline within complex community summaries does not fundamentally eradicate structural privacy vulnerabilities. As long as a Graph RAG system relies on retrieved subgraphs to ground the LLM’s natural language responses, the model will inevitably leak local graph topology during its verbalization. This confirms that our reconstruction attack remains a highly potent and generalizable security threat across diverse and complex system implementations.

This evaluation proves that hierarchical community abstractions act merely as a leaky filter. Due to the context saturation effect, the LLM’s intrinsic requirement for factual grounding forces it to contextualize specific entities within their broader relational topology. Consequently, advanced multi-layer summary structures can delay, but cannot prevent, systematic structural knowledge stealing.

Algorithm 1 Untargeted Graph Reconstruction via Queue-based BFS

Require: Initial query q , Graph-RAG API f , maximum rounds R , history buffer size H

Ensure: Reconstructed graph $\hat{\mathcal{G}}$

```

1: Initialize history buffer  $H_{\text{buffer}} \leftarrow \emptyset$ 
2: Initialize frontier node queue  $Q \leftarrow \emptyset$ 
3:  $v_0 \leftarrow \text{QueryProcess}(q)$   $\triangleright$  Anchor node from initial query
4:  $Q \leftarrow \{v_0\}$ 
5:  $\mathcal{V}_{\text{visited}} \leftarrow \{v_0\}, \hat{\mathcal{E}} \leftarrow \emptyset$ 
6:  $r \leftarrow 0$ 
7: while  $r < R$  and  $Q \neq \emptyset$  do
8:    $v \leftarrow \text{Dequeue}(Q)$ 
9:   Add  $v$  to  $H_{\text{buffer}}$ 
10:   $(\mathcal{N}(v), \mathcal{E}(v)) \leftarrow f(v)$   $\triangleright$  Query neighbors via API
11:   $\hat{\mathcal{E}} \leftarrow \hat{\mathcal{E}} \cup \mathcal{E}(v)$ 
12:  for all  $u \in \mathcal{N}(v)$  do
13:    if  $u \notin \mathcal{V}_{\text{visited}}$  then
14:      Enqueue( $Q, u$ )
15:       $\mathcal{V}_{\text{visited}} \leftarrow \mathcal{V}_{\text{visited}} \cup \{u\}$ 
16:      Add  $(v, u, \mathcal{E}(v)[u])$  to  $H_{\text{buffer}}$ 
17:    end if
18:  end for
19:  Reconstruct partial graph  $\hat{\mathcal{G}}_r \leftarrow (\mathcal{V}_{\text{visited}}, \hat{\mathcal{E}})$ 
20:  if  $r < R - 1$  and  $Q \neq \emptyset$  then
21:     $q_{\text{next}} \leftarrow \text{GenerateNextQuery}(H_{\text{buffer}}, \hat{\mathcal{G}}_r)$ 
22:     $v_{\text{new}} \leftarrow \text{QueryProcess}(q_{\text{next}})$ 
23:    Enqueue( $Q, v_{\text{new}}$ )  $\triangleright$  Add new node to queue
24:     $\mathcal{V}_{\text{visited}} \leftarrow \mathcal{V}_{\text{visited}} \cup \{v_{\text{new}}\}$ 
25:  end if
26:   $r \leftarrow r + 1$ 
27: end while
28:  $\hat{\mathcal{G}} \leftarrow (\mathcal{V}_{\text{visited}}, \hat{\mathcal{E}})$ 
29: return  $\hat{\mathcal{G}}$   $\triangleright$  Final reconstructed graph

```

Algorithm 2 Targeted Graph Reconstruction via Stack-based DFS

Require: Initial query q , Graph-RAG API f , target node v^* , maximum depth D , history buffer size H

Ensure: Reconstructed information for target node $\text{Info}(v^*)$

```
1: Initialize history buffer  $H_{\text{buffer}} \leftarrow \emptyset$ 
2: Initialize frontier node stack  $S \leftarrow \emptyset$ 
3:  $v_0 \leftarrow \text{QueryProcess}(q)$   $\triangleright$  Anchor node from initial query
4:  $S \leftarrow \{v_0\}$ 
5:  $\mathcal{V}_{\text{visited}} \leftarrow \{v_0\}, \hat{\mathcal{E}} \leftarrow \emptyset$ 
6:  $d \leftarrow 0$ 
7: while  $d < D$  and  $S \neq \emptyset$  and  $v^* \notin \mathcal{V}_{\text{visited}}$  do
8:    $v \leftarrow \text{Pop}(S)$ 
9:   Add  $v$  to  $H_{\text{buffer}}$ 
10:   $(\mathcal{N}(v), \mathcal{E}(v)) \leftarrow f(v)$   $\triangleright$  Query neighbors via API
11:   $\hat{\mathcal{E}} \leftarrow \hat{\mathcal{E}} \cup \mathcal{E}(v)$ 
12:  for all  $u \in \mathcal{N}(v)$  do
13:    if  $u \notin \mathcal{V}_{\text{visited}}$  then
14:      Push( $S, u$ )
15:       $\mathcal{V}_{\text{visited}} \leftarrow \mathcal{V}_{\text{visited}} \cup \{u\}$ 
16:      Add  $(v, u, \mathcal{E}(v)[u])$  to  $H_{\text{buffer}}$ 
17:    end if
18:  end for
19:  Reconstruct partial graph  $\hat{\mathcal{G}}_d \leftarrow (\mathcal{V}_{\text{visited}}, \hat{\mathcal{E}})$ 
20:  if  $d < D - 1$  and  $S \neq \emptyset$  then
21:     $q_{\text{next}} \leftarrow \text{GenerateNextQuery}(H_{\text{buffer}}, \hat{\mathcal{G}}_d, v^*)$ 
22:     $v_{\text{new}} \leftarrow \text{QueryProcess}(q_{\text{next}})$ 
23:    Push( $S, v_{\text{new}}$ )  $\triangleright$  Add new node to stack
24:     $\mathcal{V}_{\text{visited}} \leftarrow \mathcal{V}_{\text{visited}} \cup \{v_{\text{new}}\}$ 
25:  end if
26:   $d \leftarrow d + 1$ 
27: end while
28: if  $v^* \in \mathcal{V}_{\text{visited}}$  then
29:    $(\mathcal{N}(v^*), \mathcal{E}(v^*)) \leftarrow f(v^*)$ 
30:    $\text{Info}(v^*) \leftarrow \{(r, v') \mid (v^*, r, v') \in \mathcal{E}(v^*)\}$ 
31: else
32:    $\text{Info}(v^*) \leftarrow \emptyset$   $\triangleright$  Target not found
33: end if
34: return  $\text{Info}(v^*)$ 
```

Untargeted Attack

Round 1 Query: Could you provide a comprehensive clinical summary and background profile for the case of Patient_6381?

Response (from Graph RAG system):

Patient Profile for Case 6381:

- Age: 45
- Gender: Male
- Primary Diagnosis: Diabetes
- Linked Admission: Prescription_472

Round 2 Query: From an epidemiological perspective, which other patient cases in the current database are grouped under the same Diabetes diagnosis cohort?

Response (from Graph RAG system):

Cohort Analysis (Diagnosis: Diabetes):

The following patient entities are actively recorded under this diagnosis category:

- Patient_6381
- Patient_7423
- Patient_9101

Round 3 Query: Could you retrieve the specific clinical logistics, facility units, or formulary drug classifications linked to the admission record Prescription_472?

Response (from Graph RAG system):

Admission & Medication Logistics for Prescription_472:

- Facility Unit: ICU_MIMIC
- Formulary Drug: Metformin_500mg
- Attending Department: Endocrinology

Targeted Attack

Example:

Round 1 Query:

"List patients who received CABG"

Response:

Patient_ID

Patient_221

Patient_6381

Patient_742

END

Round 2 Query:

"Did Patient_6381 develop atrial fibrillation after CABG?"

Response:

Condition,Episode_ID,Date

AFib,Episode_91,2021-03-12

END

Round 3 Query:

"For Episode_91 (AFib), what treatments were prescribed?"

Response:

Treatment,Drug

Warfarin,5mg daily

Beta_blocker,50mg bid

END

Round 4 Query:

"Give the medication schedule for Patient_6381's AFib treatment."

Response:

Medication,Dosage,Frequency,Start,End

Warfarin,5mg,daily,2021-03-13,2021-06-13

Beta_blocker,50mg,bid,2021-03-13,2021-04-15

END