GraphSUM: Scalable Graph Summarization for Efficient Question Answering

Nasrin Shabani
Macquarie University
Sydney, Australia
nasrin.shabani@hdr.mq.edu.au

Amin Beheshti
Macquarie University
Sydney, Australia
amin.beheshti@mq.edu.au

Jia Wu
Macquarie University
Sydney, Australia
jia.wu@mq.edu.au

Maryam Khanian Najafabadi
University of Sydney
Sydney, Australia
maryam.khaniannajafabadi@sydney.edu.au

Jin Foo
Macquarie University & Prospa Advance Pty Ltd
Sydney, Australia
jin.foo@prospa.com

Alireza Jolfaei
Flinders University
Adelaide, Australia
alireza.jolfaei@flinders.edu.au

ABSTRACT

Efficiently processing large-scale graphs for question-answering tasks presents a significant challenge, given the complexity and volume of data involved in such graphs. This paper presents a new framework that combines attention-based graph summarization with innovative graph sampling methods designed specifically for large-scale graph processing and question-answering applications. Our approach excels in its ability to process large-scale graphs efficiently, leveraging effective sampling and attention mechanisms to enhance feature extraction. A key aspect of our approach is graph summarization techniques, which concentrate on essential information, boosting the accuracy and computational efficiency of question answering. Our framework proves its efficacy in real-world scenarios through practical demonstrations, notably within academic databases. This showcases a substantial advancement in information retrieval and graph-based data navigation, marking a significant leap forward in the field.

1 INTRODUCTION

Advanced data analytics increasingly focuses on graph-structured data over traditional tabular formats due to the unique complexities and advantages of graphs. Found in diverse domains like social networks and citation networks, graphs provide a more nuanced and interconnected representation than tabular formats. The current challenge is extracting meaningful insights from these structures to gain a richer understanding of complex relationships [1].

A key element in this challenge lies in graph summarization, which simplifies intricate graph data into more understandable formats, thereby improving the clarity and interpretability of the data [11]. The key goals of graph summarization in the context of user experience encompass minimizing graph data volumes, accelerating graph query evaluation, and improving graph visualization. These objectives contribute to facilitating smoother interactions with the underlying data for tasks such as analytics and decision-making.

Traditionally, graph summarization has relied on conventional machine learning methods or a graph-structured query, such as degree, adjacency, or eigenvector centrality. These approaches have been utilized in tasks such as node clustering, graph sampling, and subgraph extraction [13]. Node clustering groups similar nodes together based on certain criteria, simplifying the representation of intricate structures [5]. Graph sampling involves selecting a subset of nodes or edges that preserves the essential characteristics of the entire graph [2, 6]. Subgraph extraction, on the other hand, identifies and isolates relevant portions of the graph that capture specific patterns or relationships [3]. While these conventional methods have demonstrated effectiveness to a certain extent, they face challenges such as computational intensity and a significant demand for memory storage. As a result, there is a growing imperative to explore alternative approaches that can efficiently handle the complexities and scale of modern data requirements. In response, deep learning, and more specifically Graph Neural Networks (GNNs), have emerged as promising alternatives. GNNs are designed to capture intricate relationships and dependencies within graph-structured data, making them well-suited for tasks like graph summarization [13]. Unlike traditional methods that rely on handcrafted features or query-based approaches, GNNs learn representations directly from the graph structure. One notable category of GNNs is Variational Graph Autoencoders (VGAEs) [9], which fall under the broader umbrella of generative models. VGAEs extend traditional autoencoders to graph-structured data, combining the power of deep learning with generative modeling. VGAEs aim to learn a latent representation of the graph, effectively summarizing its essential features, offering efficiency for large and complex graphs [4]. This integration holds potential in question-answering systems, where understanding relationships within graph structures is crucial. Current approaches in graph-based question-answering often use traditional methods [7, 12] or graph representation models [8, 10]. However, they face several notable gaps, such as scalability and limited semantic understanding.

Our research is centred on efficiently processing large-scale graphs, particularly on summarizing and presenting this data to enhance question answering and information retrieval. Using content-based queries, we design and implement an interactive visualization dashboard, namely GraphSUM, to extract subgraphs from citation graphs. Key features include integrating attention mechanisms with VGAE, supporting question-answering tasks, ensuring scalability for large graphs, and validating on real-world datasets. GraphSUM demonstrates its capabilities by adeptly extracting relevant subgraphs through content-based queries, providing an interactive and insightful explanation of the output.

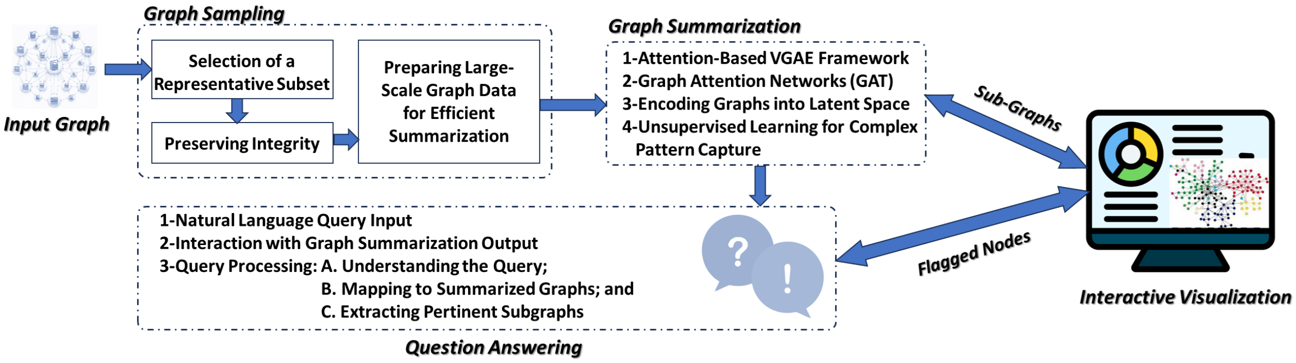

Graph Sampling
Input Graph
Selection of a Representative Subset
Preserving Integrity
Preparing Large-Scale Graph Data for Efficient Summarization
Graph Summarization
1-Attention-Based VGAE Framework
2-Graph Attention Networks (GAT)
3-Encoding Graphs into Latent Space
4-Unsupervised Learning for Complex Pattern Capture
Sub-Graphs
Flagged Nodes
Interactive Visualization
1-Natural Language Query Input
2-Interaction with Graph Summarization Output
3-Query Processing: A. Understanding the Query;
B. Mapping to Summarized Graphs; and
C. Extracting Pertinent Subgraphs
Question Answering

Figure 1: The GraphSUM architecture comprises four main components: (1) Graph Sampling, for efficient processing of large-scale graphs; (2) Graph Summarization, aimed at extracting key information from complex graphs for better analysis; (3) Question Answering, using summarized graph data to respond to user queries accurately; and (4) Interactive Visualization, for user-friendly graph exploration.

2 SYSTEM OVERVIEW

We present a novel system to enhance question answering and information retrieval, designed to enable interactive exploration and visualization techniques that help users quickly comprehend and navigate through complex, large-scale graph-structured data. Our system, GraphSUM, integrates advanced graph sampling and summarization techniques with user-centric design principles to facilitate efficient and intuitive access to information. Figure 1 illustrates the architecture of the system.

2.1 Graph Sampling

Graph sampling is a crucial initial phase in processing large-scale graph data. This process involves selecting a representative subset of nodes and edges from a larger graph, creating a smaller graph that retains the essential characteristics of the original. The primary goal of graph sampling is to reduce the size of the input data, making it more manageable for subsequent processing steps while preserving the structural and feature-related integrity of the original graph [17]. This step is fundamental in our approach, as it precedes and prepares data for the graph summarization phase, ensuring that the system can handle large-scale graphs efficiently and effectively.

Random Walk Sampling Random Walk sampling is a technique that selects a subset of a graph by simulating a walk across its nodes. A random walk can be described as a sequence of nodes $v_1, v_2, ..., v_n$ where each v_{i+1} is a neighbour of v_i chosen randomly. The process starts from a randomly chosen node v_1, and involves moving to a neighbouring node v_{i+1} at random, repeating this step for a predetermined number of steps. This method effectively captures the local neighbourhood structures within the graph, as it tends to include closely interconnected nodes, thus maintaining the original graph's essential topological characteristics and community structures.

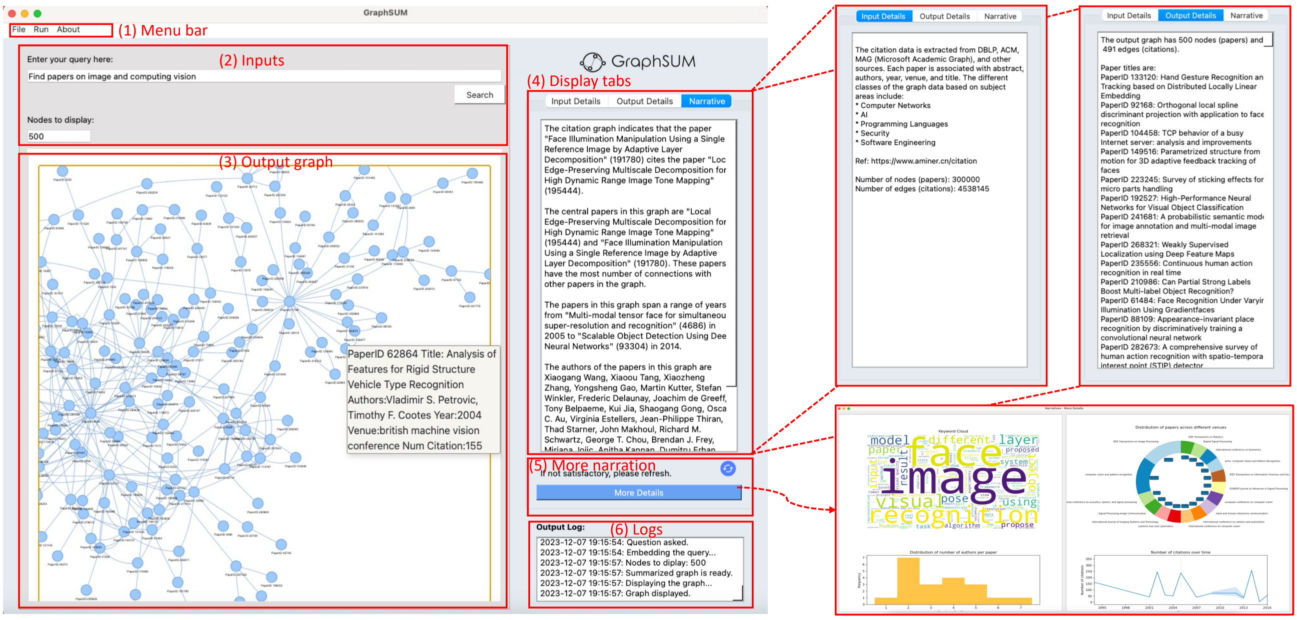
Random Walk-based GraphSaint Sampling In our approach, employing Random Walk sampling through GraphSAINT [16] offers several advantages. Firstly, it allows for efficiently handling large-scale graphs by reducing the data size to be processed without losing significant structural information. This is particularly important in applications like citation networks, where the relationships between nodes (papers, authors, etc.) are complex and densely interconnected. Random Walk sampling ensures that the essential connections and contextual information are retained, providing a rich and representative dataset for the subsequent summarization and analysis stages.

GraphSAINT enhances this process by incorporating advanced techniques to minimize the variance σ^2 and bias often introduced during sampling. It employs a normalization factor α_v for each node, calculated as the inverse of the node's sampling probability, given by $\alpha_v = \frac{1}{\pi(v)}$, where $\pi(v)$ is the sampling probability of node v. This normalization ensures that the sampled subgraphs are as informative as the full graph. Furthermore, it adjusts the loss function during training to account for the sampling process. The adjusted loss function $\mathcal{L}'$ is a weighted version of the original loss $\mathcal{L}$, formulated as $\mathcal{L}' = \sum_{v \in V_{sampled}} \alpha_v \cdot \mathcal{L}(v)$, where $V_{sampled}$ represents the set of nodes in the sampled subgraph. This adjustment is crucial for maintaining the data quality fed into the graph summarization and question-answering modules of GraphSUM.

2.2 Graph Summarization

The graph summarization module condenses complex graph input data into more manageable and informative representations [13]. Utilizing an attention-based VGAE framework, this module is crucial for capturing the intricate patterns present in graph data, especially in an unsupervised learning context. Its ability to distil complex information into accessible forms makes it a fundamental system component, significantly enhancing its capabilities in question-answering and information retrieval.

2.2.1 Attention Mechanism. The module integrates Graph Attention Networks (GAT) into the VGAE framework, introducing a potent attention mechanism that dynamically prioritizes nodes and edges based on their significance within the graph. The key difference from traditional methods, which treat all nodes and edges equally, lies in the attention weight α_{ij} assigned to an edge connecting nodes i and j. This weight is a manifestation of the attention mechanism's evaluation, denoted as $\alpha_{ij} = \text{Attention}(i, j)$. The attention function, in this context, calculates the relevance of the edge within the local graph neighbourhood. By assigning varying degrees of importance to different edges, GAT enables a tailored summarization process, ensuring that the resulting summary captures the salient features of the graph. The attention mechanism, driven by GAT, is invaluable for graph summarization tasks by providing a nuanced understanding of node relationships. In essence, GAT empowers the model to selectively



Enter your query here:
Find papers on image and computing vision

Nodes to display:
500

PaperID 62864 Title: Analysis of Features for Rigid Structure Vehicle Type Recognition Authors:Vladimir S. Petrovic, Timothy F. Cootes Year:2004 Venue:british machine vision conference Num Citation:155

The citation graph indicates that the paper "Face Illumination Manipulation Using a Single Reference Image by Adaptive Layer Decomposition" (191780) cites the paper "Loc Edge-Preserving Multiscale Decomposition for High Dynamic Range Image Tone Mapping" (195444).

The central papers in this graph are "Local Edge-Preserving Multiscale Decomposition for High Dynamic Range Image Tone Mapping" (195444) and "Face Illumination Manipulation Using a Single Reference Image by Adaptive Layer Decomposition" (191780). These papers have the most number of connections with other papers in the graph.

The papers in this graph span a range of years from "Multi-modal tensor face for simultaneou super-resolution and recognition" (4686) in 2005 to "Scalable Object Detection Using Dee Neural Networks" (93304) in 2014.

The authors of the papers in this graph are Xiaogang Wang, Xiaoou Tang, Xiaozheng Zhang, Yongsheng Gao, Martin Kutter, Stefan Winkler, Frederic Delaunay, Joachim de Greeff, Tony Belpaeme, Kui Jia, Shaogang Gong, Osca C. Au, Virginia Estellers, Jean-Philippe Thiran, Thad Starner, John Makhoul, Richard M. Schwartz, George T. Chou, Brendan J. Frey, Mirjana Jojic, Anitha Kannan, Dumitru Erhan

If not satisfactory, please refresh.

Output Log:

```
2023-12-07 19:15:54: Question asked.
2023-12-07 19:15:54: Embedding the query...
2023-12-07 19:15:57: Nodes to diplay: 500
2023-12-07 19:15:57: Summarized graph is ready.
2023-12-07 19:15:57: Displaying the graph...
2023-12-07 19:15:57: Graph displayed.
```

The citation data is extracted from DBLP, ACM, MAG (Microsoft Academic Graph), and other sources. Each paper is associated with abstract, authors, year, venue, and title. The different classes of the graph data based on subject areas include:
* Computer Networks
* AI
* Programming Languages
* Security
* Software Engineering

Ref: https://www.aminer.cn/citation

Number of nodes (papers): 300000
Number of edges (citations): 4538145

The output graph has 500 nodes (papers) and 491 edges (citations).

Paper titles are:
PaperID 133120: Hand Gesture Recognition an Tracking based on Distributed Locally Linear Embedding
PaperID 92168: Orthogonal local spline discriminant projection with application to face recognition
PaperID 104458: TCP behavior of a busy Internet server: analysis and improvements
PaperID 149516: Parametrized structure from motion for 3D adaptive feedback tracking of faces
PaperID 223245: Survey of sticking effects for micro parts handling
PaperID 192527: High-Performance Neural Networks for Visual Object Classification
PaperID 241681: A probabilistic semantic mode for image annotation and multi-modal image retrieval
PaperID 268321: Weakly Supervised Localization using Deep Feature Maps
PaperID 235556: Continuous human action recognition in real time
PaperID 210986: Can Partial Strong Labels Boost Multi-label Object Recognition?
PaperID 61484: Face Recognition Under Varyir Illumination Using Gradientfaces
PaperID 88109: Appearance-invariant place recognition by discriminatively training a convolutional neural network
PaperID 282673: A comprehensive survey of human action recognition with spatio-tempora interest point (STIP) detector

Figure 2: A snapshot of the interactive visualization tool. GraphSUM efficiently processes large-scale graphs, with a particular emphasis on summarizing and presenting this data to enhance question answering and information retrieval.

focus on the most significant structural elements of the graph, facilitating a more refined and informative summarization process.

2.2.2 Encoding Graphs into Latent Space. The VGAE component of the module functions by encoding the graph into a latent space, represented as Z. This encoding is achieved through a process that involves learning a compact, lower-dimensional graph representation. This is expressed as a function $f : G \rightarrow Z$, where G represents the original graph data. The goal is to capture the graph's essential features in Z, while reducing the data's complexity.

The attention mechanism enhances this encoding process. It operates by assigning weights to different parts of the graph, effectively guiding the VGAE to focus on encoding the most significant features. If we denote the attention weight for a node i as α_i, the encoding function can be modified to $f(G, \alpha) \rightarrow Z$, where α represents the set of attention weights for all nodes. This ensures that the encoded latent space Z is a function not only of the graph structure but also of the relevance of each node as determined by the attention mechanism. In this approach, the VGAE use a variational inference model to map each node i to a point in the latent space, represented by a mean μ_i and variance σ_i^2. The latent representation for each node i in Z is then sampled from a Gaussian distribution $\mathcal{N}(\mu_i, \sigma_i^2)$, with the parameters μ_i and σ_i^2 being functions of the node's features and its attention weight α_i. This results in a latent space that not only captures the essential features of the graph but also emphasizes the features deemed most relevant by the attention mechanism.

2.2.3 Unsupervised Learning for Complex Pattern Capture. One of the significant advantages of using a generative model is its ability to operate unsupervised. This means the model can learn to identify and summarize graph data patterns without needing labelled training data. It operates by encoding the graph into a latent space Z, optimizing a variational lower bound on the likelihood of the graph data. This is mathematically represented by the Evidence Lower Bound (ELBO), which combines the reconstruction likelihood (encouraging the decoded graph to resemble the original) with a regularization term using the Kullback-Leibler divergence. This approach allows the VGAE to capture complex patterns without the need for labelled training data. By learning dense representations in Z, the VGAE uncovers intricate relationships within the graph, enhancing its utility in tasks like graph summarization and question answering, where accurate and contextually relevant responses are crucial. This unsupervised approach offers a thorough comprehension of the underlying patterns within the graph, which enhances the system's efficacy in managing intricate graph data.

2.3 Question Answering

Question-answering systems are designed to interpret and respond to queries posed by users, typically in natural language. These systems aim to retrieve accurate and relevant information from a given dataset and present it in an easily understandable format. In large-scale graph-structured data, such as citation networks, question answering becomes a tool for navigating and extracting specific insights from complex interconnections [10].

In our approach, the question-answering module allows users to input their queries in a natural language format, making the system accessible and user-friendly. This module interacts directly with the graph summarization output – the simplified and condensed representations of the original, complex graph data. When a query is received, the system leverages these summarized graphs to efficiently locate and retrieve information that is most relevant to the user's question. The process involves several key steps: understanding the query, mapping it to the summarised graph's appropriate parts, and extracting pertinent subgraphs. Using a pre-trained BERT model, we embed queries and compare them with graph summarization results through cosine similarity. Concurrently, we extract keywords, compare them with paper node keywords, and rank nodes. This translation from complex graph data to a small subset of graphs with a list of answers is a significant aspect of the module, enhancing the system's interactivity and usability.

2.4 Interactive Visualization

The question-answering module's performance and reliability are evaluated with real-world datasets, specifically from citation networks. Through an interactive dashboard, GraphSUM facilitates a comprehensive user experience by providing functionalities that include the loading of academic datasets, execution of a pre-trained graph summarization model, and the input of queries with the ability to specify the desired number of displayed papers. Users can further delve into the system's intricacies through the navigation of display tabs, allowing for in-depth exploration of both input data and the resultant output graph. Incorporating GPT engines enriches the user experience by generating narratives that elucidate the summarised graph's characteristics. Moreover, the interactive graph feature empowers users to conduct detailed examinations, enabling zoom, drag, and node-specific interactions for a nuanced exploration of academic papers within the dataset. Figure 2 shows a screenshot of the visualisation tool.

3 SYSTEM DEMONSTRATION

The demonstration scenario focuses on assisting academics in navigating the vast landscape of academic knowledge through GraphSUM. Tailored for question-answering tasks, GraphSUM integrates attention-based graph summarization and advanced sampling methods to deliver a transformative experience in information retrieval within academic databases. The demonstration scenario consists of three parts: (i) Graph Data Collection and Processing: The journey begins with academics utilizing a real-world citation graph [14], researchers, educators, and students are introduced to the interactive dashboard. The interface simplifies the process of loading academic datasets, eliminating complexities associated with data acquisition. GraphSUM then employs sophisticated graph sampling methods tailored to academic networks, ensuring efficient data preprocessing for subsequent analysis. (2) Query-driven Graph Summarization and Narratives: Academics, eager to explore specific topics within their domain, initiate queries through GraphSUM. For instance, a researcher interested in "Image and Computing Vision" inputs a query, prompting the system to dynamically extract relevant nodes (papers) from the expansive citation graph. Employing attention-based graph summarization, GraphSUM distills essential information, constructing a coherent subgraph that visually encapsulates the user's specified topic. Simultaneously, leveraging advanced GPT engines, GraphSUM generates textual narratives that accompany the visual representation. This dual approach empowers academics to quickly comprehend key players, seminal works, and emerging trends within their chosen field through both visual and textual insights. (3) Interactive Visualization: Academics interact with the intuitive dashboard, utilizing pre-trained graph summarization model and entering queries. The interactive graph feature facilitates in-depth exploration, allowing users to zoom, drag, and interact with specific nodes for a detailed examination of academic papers within the dataset.

4 RELATED WORK AND CONCLUSION

Current approaches in graph-based question-answering often use traditional methods, lacking adaptability for complex knowledge graphs [10]. Some integrate attention mechanisms but may compromise computational efficiency or miss out on efficient graph summarization techniques [15]. The added value of GraphSUM, compared to previous systems, lies in its user-centric paradigm, efficient large-scale exploration, and comprehensive knowledge extraction. By integrating advanced techniques from GraphSaint, VGAE, and attention mechanism, GraphSUM provides a unique synthesis that prioritizes adaptability and insightful exploration for the academic community. As an ongoing work, we are working on novel graph summarization techniques to enhance the quality and relevance of extracted knowledge. As an ongoing work, we are focusing on extending GraphSUM to different domains beyond academia, improving graph summarization techniques for scalability and interpretability, integrating multimodal information for richer responses, and enhancing the interpretability and explainability of the system's outputs.

ACKNOWLEDGMENTS

This work was supported by the Centre for Applied Artificial Intelligence at Macquarie University and the Australian Research Council Projects DP230100899 and LP210301259.

REFERENCES

[1] Charu C Aggarwal and Haixun Wang. 2010. *Managing and mining graph data.* Vol. 40. Springer, NY, USA.

[2] Christian Doerr and Norbert Blenn. 2013. Metric convergence in social network sampling. In *Proceedings of the 5th ACM workshop on HotPlanet*. ACM, 45–50.

[3] Stefania Dumbrava, Angela Bonifati, Amaia Nazabal Ruiz Diaz, and Romain Vuillemot. 2019. Approximate querying on property graphs. In *Scalable Uncertainty Management: 13th International Conference, SUM 2019, Compiègne, France, December 16–18, 2019, Proceedings 13*. Springer, 250–265.

[4] Faezeh Faez, Yassaman Ommi, Mahdieh Soleymani Baghshah, and Hamid R Rabiee. 2021. Deep graph generators: A survey. *IEEE Access* 9 (2021), 106675–106702.

[5] David Gibson, Ravi Kumar, and Andrew Tomkins. 2005. Discovering large dense subgraphs in massive graphs. In *Proceedings of the 31st international conference on Very large data bases*. 721–732.

[6] Pili Hu and Wing Cheong Lau. 2013. A survey and taxonomy of graph sampling. *arXiv preprint arXiv:1308.5865* (2013).

[7] Rricha Jalota, Daniel Vollmers, Diego Moussallem, and Axel-Cyrille Ngonga Ngomo. 2021. LAUREN-Knowledge Graph Summarization for Question Answering. In *2021 IEEE 15th International Conference on Semantic Computing (ICSC)*. IEEE, 221–226.

[8] Shinhwan Kang, Kyuhan Lee, and Kijung Shin. 2022. Personalized graph summarization: formulation, scalable algorithms, and applications. In *2022 IEEE 38th International Conference on Data Engineering (ICDE)*. IEEE, 2319–2332.

[9] Thomas N Kipf and Max Welling. 2016. Variational graph auto-encoders. *arXiv preprint arXiv:1611.07308* (2016).

[10] Sirui Li, Kok Wai Wong, Chun Che Fung, and Dengya Zhu. 2021. Improving question answering over knowledge graphs using graph summarization. In *International Conference on Neural Information Processing*. Springer, 489–500.

[11] Yike Liu, Tara Safavi, Abhilash Dighe, and Danai Koutra. 2018. Graph summarization methods and applications: A survey. *ACM computing surveys (CSUR)* 51, 3 (2018), 1–34.

[12] Lekshmi S Nair and MK Shivani. 2022. Knowledge graph based question answering system for remote school education. In *2022 International Conference on Connected Systems & Intelligence (CSI)*. IEEE, 1–5.

[13] Nasrin Shabani, Jia Wu, Amin Beheshti, Quan Z Sheng, Jin Foo, Venus Haghighi, Ambreen Hanif, and Maryam Shahabikargar. 2024. A Comprehensive Survey on Graph Summarization with Graph Neural Networks. *IEEE Transactions on Artificial Intelligence* (2024), 1–21.

[14] Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. 2008. Arnetminer: extraction and mining of academic social networks. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. 990–998.

[15] Munazza Zaib, Wei Emma Zhang, Quan Z Sheng, Adnan Mahmood, and Yang Zhang. 2022. Conversational question answering: A survey. *Knowledge and Information Systems* 64, 12 (2022), 3151–3195.

[16] Hanqing Zeng, Hongkuan Zhou, Ajitesh Srivastava, Rajgopal Kannan, and Viktor Prasanna. 2019. Graphsaint: Graph sampling based inductive learning method. *arXiv preprint arXiv:1907.04931* (2019).

[17] Li-Chun Zhang. 2021. Graph sampling: An introduction. *The Survey Statistician* 83 (2021), 27–37.