# HeteroCS: A Heterogeneous Community Search System With Semantic Explanation

Weibin Cai
Hangzhou City University
Hangzhou, China
c549971521@gmail.com

Fanwei Zhu[*]
Hangzhou City University
Hangzhou, China
zhufanwei@zju.edu.cn

Zemin Liu
National University of Singapore
Singapore, Singapore
zeminliu@nus.edu.sg

Minghui Wu
Hangzhou City University
Hangzhou, China
mhwu@zucc.edu.cn

## ABSTRACT

Community search, which looks for query-dependent communities in a graph, is an important task in graph analysis. Existing community search studies address the problem by finding a densely-connected subgraph containing the query. However, many real-world networks are heterogeneous with rich semantics. Queries in heterogeneous networks generally involve in multiple communities with different semantic connections, while returning a single community with mixed semantics has limited applications. In this paper, we revisit the community search problem on heterogeneous networks and introduce a novel paradigm of heterogeneous community search and ranking. We propose to automatically discover the query semantics to enable the search of different semantic communities and develop a comprehensive community evaluation model to support the ranking of results. We build HeteroCS, a heterogeneous community search system with semantic explanation, upon our semantic community model, and deploy it on two real-world graphs. We present a demonstration case to illustrate the novelty and effectiveness of the system.

## CCS CONCEPTS

• **Computing methodologies** → **Artificial intelligence**; • **Information systems** → *Information retrieval*.

## KEYWORDS

community search, query semantics, heterogeneous community

**ACM Reference Format:**
Weibin Cai, Fanwei Zhu, Zemin Liu, and Minghui Wu. 2023. HeteroCS: A Heterogeneous Community Search System With Semantic Explanation. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23), July 23–27, 2023, Taipei, Taiwan.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3539618.3591812

## 1 INTRODUCTION

Communities naturally exist in numerous real-world graphs such as collaboration networks, online social networks, *etc.* Traditionally, community-level analysis in graphs mainly focuses on detecting community structures within the entire graph, which has limited the application in query-dependent scenarios. Recently, community search [17, 2, 10] (CS for short), which aims to find the densely connected subgraph containing the query nodes in the graph, has attracted a surge of research attention. Most CS works focus on querying the communities on homogeneous networks [1, 4, 8, 14]. However, as many real-world graphs are heterogeneous [20, 16], ignoring the heterogeneity, *i.e.*, types of nodes and links, would miss the valuable semantic information for community mining. Recent advances in CS [5, 12, 11] usually attempt to model the relationship between nodes in heterogeneous networks by the meta-paths connecting them [5, 13, 12] or the user-specified relational constraints [11]. Nevertheless, meta-paths based methods are designed to find a homogeneous community of nodes with the same type rather than to discover heterogeneous communities involving diverse types of nodes; relational constraint-based methods require the users to have a good understanding of the community schema thus to define a proper set of constraints, otherwise, it may fail to find any community without any assistance. Moreover, all existing CS works only return one community containing the query; whereas in real scenarios, a query can be involved in multiple communities with different semantics, especially when the network schema is complex and the query intent is diverse.

In this paper, we introduce a new problem of *semantic community search* over heterogeneous graphs that searches for a list of heterogeneous communities to provide better community exploration from different semantic aspects, and develop **HeteroCS**, a heterogeneous community search system with semantic explanation and evaluation.

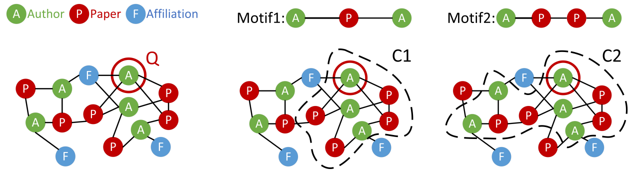
**Problem.** To tackle the heterogeneity of real-world graphs and the diversity of query semantics, we define semantic community search as finding a ranked list of query-relevant communities with different semantics.

[*] Corresponding author.

*SIGIR '23, July 23–27, 2023, Taipei, Taiwan*

ACM ISBN 978-1-4503-9408-6/23/07…$15.00
https://doi.org/10.1145/3539618.3591812

As shown in Fig. 1, given an HIN with different types of nodes and edges, for a specific query (*e.g.*, Q in Fig. 1(a)), we detect a set of communities related to Q (*e.g.*, C1, C2) and explicitly profile them by the *motifs* (*i.e.*, meta-paths or meta-structures) frequently appeared in the communities. For example, C1 consists of many connections whose meta-path is *Author-Paper-Author*, which reveals that community C1 is the collaboration network of Q, formed by the co-authors of the query node and their collaborations. To support the semantic community search and ranking, we need to tackle three main tasks: (1) understand query intent or query semantics, (2) discover query-relevant communities, and (3) evaluate the candidate communities.



(a) Toy graph with query Q. (b) Community C1. (c) Community C2.

**Figure 1: An illustration example of semantic community search. Given a query Q, there are two qualified communities with different semantics: C1 is formed by co-authorship and C2 is formed by paper citations.**

For the first task, we construct a query subgraph consisting of different types of links from the query and discover the frequent motifs in the subgraph to model the query semantics. With query semantics explicitly modeled, we next train a GCN model with contrastive learning technique to generate the community-aware representation of each node such that nodes that belong to the query-relevant semantic community would have more similar representations as the query compared to the other nodes. Based on the node representations, an iterative community expansion algorithm is developed to expand the semantic community from the query. We comprehensively evaluate candidate communities with different semantics, in terms of their query relevance and structural cohesiveness. Specifically, query relevance measures how a community is related to the query, both geometrically and semantically, and structural cohesiveness evaluates if it is a good community with dense inter-connections. Communities with different semantics are returned in a ranked order along with their explanations.

**System.** We build a system HeteroCS to demonstrate the effectiveness and interpretability of the semantic community search. It consists of three major function modules: *query semantic discovery*, *semantic-aware community search* and *community evaluation*. The query semantic discovery module constructs a query-centered subgraph in which the query semantics is discovered to guide the search of heterogeneous communities. The search module then takes the query nodes and the corresponding semantics to generates a ranked list of communities. The structure of each community, along with its semantic explanation is visualized, and the ranking scores returned by the evaluation module such as query relevance, cohesiveness are also exhibited in the system. The detailed design of the system is elaborated in Sect. 3.

**Contributions.** The main contributions of this paper are threefold: (1) We revisit the CS problem over heterogeneous graphs and propose a new paradigm of semantic community search and ranking with explanations. (2) We develop a community-aware GCN model, an iterative community expansion algorithm and a comprehensive evaluation model to support effective community search and ranking. (3) We build a demo system that implements the semantic discovery, semantic community search and explanation, and present a case study to illustrate the usability and effectiveness of the system.

## 2 RELATED WORK

Different from community detection that identifies all communities in a network, community search can be regarded as query-dependent community discovery. Most of the current community search works focus on finding a dense subgraph containing the query on homogeneous networks. Different structure metrics such as $k$-core, $k$-truss, $k$-clique are adopted to measure the cohesiveness of a community. For instance, Sozio *et al.* [17] propose a core-based community search model that finds a largest $k$-core containing the query as the community. Huang *et al.* [9] present a truss-based community search method that ensures each edge in the community to be contained within at least $k$-2 triangles. Recently, a growing number of researchers have shown interests in community search problem over heterogeneous networks (or CSH problem). As the first CSH work, Fang *et al.* [5] propose to use a meta-path $P$ to define the connectivity of different nodes in an HIN, and extend the $k$-core metric to $(k, \mathcal{P})$-core to measure the cohesiveness of a community. Jiang *et al.* [13] further improve the CSH model to automatically generate the maximal set of qualified meta-paths and search for the community that shares these meta-paths. However, these two works are designed to find homogeneous communities in an HIN, that is, all nodes in a community are with the same type. To search for the heterogeneous community, Jian *et al.* [11] propose a relational community model RCS that defines a community upon a set of user-specified relational constraints. Nodes satisfy all the relational constraints form a relational community and the minimal community containing the query is returned. Although RCS can handle personalized community requirements, it may fail to find any community if the constraints are not properly specified.

## 3 SYSTEM DESCRIPTION

The overall framework of our HeteroCS system is shown in Fig. 2. It provides an interactive search interface– initially user selects a dataset and the nodes to query, which are then processed by the backend *Query Semantic Discover Module* to obtain the query subgraph and query semantics. The query subgraph and semantics are visualized and user can further choose certain semantic schema to query. After receiving the query semantic from the user, the *Semantic-aware Community Search Module* searches for the corresponding semantic community and the *Community Evaluation Module* comprehensively evaluates the community, and then the community as well as its goodness scores and semantic explanations are returned to the user. The detailed design of the system is described in the following subsections.

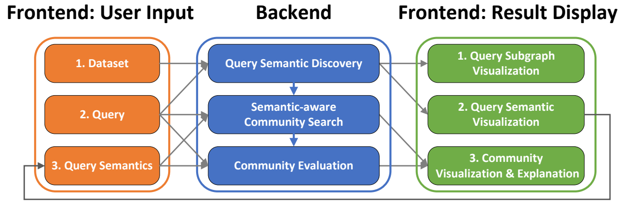

Frontend: User Input
Backend
Frontend: Result Display
1. Dataset
2. Query
3. Query Semantics
Query Semantic Discovery
Semantic-aware Community Search
Community Evaluation
1. Query Subgraph Visualization
2. Query Semantic Visualization
3. Community Visualization & Explanation

**Figure 2: System Architecture**

## 3.1 Frontend

The frontend of the HeteroCS system consists of two parts: User Input and Result Display. User Input receives user inputs and sends them to the backend for processing. Result Display renders the results passed from the backend on the interface. There are two types of inputs required in HeteroCS system:

(1) *Dataset and Query.* User can select a dataset and enters a query to search for the query-dependent community in the dataset. The request is sent to the backend to retrieve the query-relevant subgraph (or query subgraph for short), as well as a list of possible query semantics. The frontend renders the query subgraph and the network schema to the interface, and displays the query semantics as a table object where each row representing a specific semantic pattern (*i.e.*, motif) and the proportion of its instances in the subgraph.

(2) *Semantics.* To search for semantic communities, the user may also need to select at least one semantic pattern in the table. The selected semantic patterns are then sent to the backend which invokes the Semantic-aware Community Search module to be executed. Corresponding semantic communities are retrieved and rendered into a collapsible component, where each item corresponding to a community includes the visualization of its structure, its evaluation scores, and the instances of its semantic pattern.

## 3.2 Query Semantic Discovery

Unlike the traditional free-text search where the query semantics can be easily inferred from the keywords, in heterogeneous community search, the query contains only nodes whose semantics is implicit in the complex link structure of the network. For example, two *Author* nodes connected by a *Paper* node (*i.e.*, *Author-Paper-Author*) in an academic network indicates the semantics of "co-authorship". Usually, meta-paths [19, 7] or meta-structures [22, 6, 21, 15] are used to represent such semantic connections in heterogeneous networks, which we also refer to as *motifs* in this paper.

As heterogeneous networks may contain diverse and complex motifs, it is quite challenging for users to manually define the query semantics to initialize the search process. To solve the problem, the Query Semantic Discovery module is designed to automatically discover and rank the possible query semantic pattern based on a query-relevant subgraph. It consists of the following two steps:

(1) *Query Subgraph Construction.* Intuitively, the semantic information of a node often exists in its neighborhood subgraph. Thus, when the system receives a query, HeteroCS first obtains the neighborhood subgraph of the query node using breath-first search method. To avoid including the remote and less relevant nodes, the maximal depth of tours in the query subgraph is set as 3.

(2) *Query Semantic Mining.* After obtaining the query subgraph, frequent motifs in the subgraph are adopted to represent the query semantics. Specifically, HeteroCS first uses the Grami algorithm[3] to obtain a set of frequently appeared motifs in the heterogeneous graph. For each motif, it mines its instances in the query subgraph and calculates its ratio among all motif instances. Top-ranked motifs with higher instance ratios are returned to represent query semantics.

## 3.3 Semantic-aware Community Search

To facilitate the semantic community search, HeteroCS first obtains the embedding of each node in the query subgraph through a community-aware GCN model, and then uses these embeddings to iteratively expand the target community.

### 3.3.1 *GNN Based Representation Learning.*

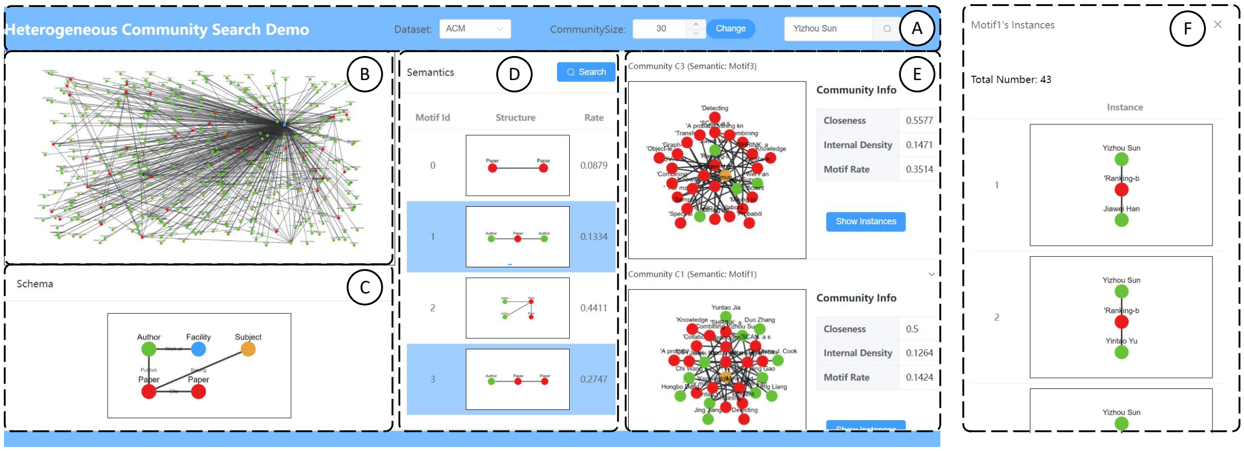
In order to obtain a query-dependent, semantic-relevant, and structure-cohesive community, we expect the node embedding to capture its relevance to the query nodes, query semantic and its structure information. To achieve this, we propose three loss functions from different perspectives, and train a community-aware GCN model that jointly optimize them. Contrastive learning technique [19] is applied in our model training as there is no labeled data for supervision.

*First,* to ensure nodes in the community are closely related to the query, we sample the positive nodes and negative nodes based on their personalized PageRank value (PPV) [23] to the query. Nodes with high PPV are good community members and thus their embeddings should be more similar to the embedding of the query nodes, while the embeddings of negative samples should be distinct from the query embedding. *Second,* to achieve a community with consistent semantic explanation, nodes in the community should be able to form as many instances of the query motif as possible. Thus, we use the motif instances in the network as positive samples and a random node set as negative samples. A trainable parameter is set for each motif to ensure the embeddings of the positive samples are close to the motif embedding and the embedding of the negative samples is far away from the motif embedding. *Third,* to make sure the community is structure-cohesive, we expect its members to be high-degree nodes to form more interconnections. Thus, we use the node degree for positive and negative sampling. Embeddings of the positive samples are optimized to be similar to query embedding, while embeddings of the negative sample are dissimilar.

Therefore, we can feed the query subgraph into the above GCN model to generate the community-aware representation of each node in the subgraph, which will be utilized to obtain the semantic community by an iterative community expansion algorithm.

### 3.3.2 *Iterative Community Expansion.*

When obtaining the embedding of each node in the query subgraph, the iterative community expansion algorithm will expand a community from the query node by iteratively including candidate nodes with the highest *priority scores* calculated over their embeddings. Specifically, in each iteration, the one-hop neighbors of current community members are selected as candidates, and the priority score of each candidate is calculated as the product of the similarity of its embedding to the



Dataset: ACM

CommunitySize: 30

Yizhou Sun

Semantics

| Motif Id | Structure | Rate |
|---|---|---|
| 0 | Paper – Paper | 0.0879 |
| 1 | Author – Paper – Author | 0.1334 |
| 2 | Author, Author, Paper, Paper | 0.4411 |
| 3 | Author – Paper – Paper | 0.2747 |

Community C3 (Semantic: Motif3)

Community Info

| | |
|---|---|
| Closeness | 0.5577 |
| Internal Density | 0.1471 |
| Motif Rate | 0.3514 |

Community C1 (Semantic: Motif1)

Community Info

| | |
|---|---|
| Closeness | 0.5 |
| Internal Density | 0.1264 |
| Motif Rate | 0.1424 |

Motif1's Instances

Total Number: 43

| | Instance |
|---|---|
| 1 | Yizhou Sun – 'Ranking-b – Jiawei Han |
| 2 | Yizhou Sun – 'Ranking-b – Yintao Yu |
| | Yizhou Sun |

**Figure 3: HeteroCS demonstration system**

query embedding, the motif embedding, and the current community embedding which is the mean of its members' embeddings. The candidate with the highest score is added to expand the current community. The iterative expansion process is terminated when the current community reaches a pre-defined size.

*3.3.3 Community Evaluation.* As mentioned earlier, the query in a heterogeneous network can have different semantics, and thus it corresponds to different semantic communities. The community evaluation module comprehensively evaluates each community from three aspects: *query relevance*, *semantic consistency* and *structure cohesiveness*, and returns the evaluation scores to the user as the *support* of search results. Specifically, the reciprocal of the sum of the shortest distances from all nodes in the community to the query node (*i.e.*, query centrality) is calculated to measure the query relevance of a community; the proportion of the query motif instances in the community (*i.e.*, motif rate) is used to measure the semantic consistency, and the relative ratio of the number of edges to the number of nodes in the community (*i.e.*, internal density) is adopted to measure the structure cohesiveness. All semantic communities of the query are evaluated and returned to the user in a ranked order, indicating their importance or relevance to the query.

## 4 DEMONSTRATION

In this section, we first demonstrate the functionalities of HeteroCS and then analyze the effectiveness of our system in a real search scenario. The video recording of the demonstration can be found at: https://github.com/ACECWB/HeteroCS.

*Demonstration scenario.* We demonstrate HeteroCS on the ACM dataset [18]. As shown in Fig. 3, we begin by selecting the dataset from the drop-down menu in Panel A. The graph structure of the dataset and the network schema are respectively displayed in Panel B and Panel C, providing users with a visual perception of the structure and semantics of the dataset. Now suppose we want to explore the communities of Professor Yizhou Sun, we enter the keyword "Yizhou Sun" in the search box and click the query button in Panel A. Different semantics (*e.g.*, *Paper-Paper*, *Author-Paper-Author)* associated with the query, along with their corresponding proportions, are shown in Panel D. We choose two types of semantics (motif 1 and motif 3), specify the community size as 30, and click the query button in Panel D to query. The corresponding semantic communities are sorted and visualized in Panel E. The evaluation scores of these communities are also displayed. For each community, we can further click the "Show Instances" button to see all the instances of the corresponding motif (*e.g.*, there are 43 instances of Motif 3 *Author-Paper-Author* in community C3), as shown in Panel F.

*Effectiveness study.* In the above search scenario, community C3 (motif rate=0.27) is ranked higher than community C1 (motif rate=0.13) according to their motif rate in the query subgraph, indicating that C3 is the primary community of the query compared to C2. By examining the scores of the other evaluation metrics (*e.g.*, closeness and internal density), we can clearly see that community C3 is indeed superior than community C1, which verifies the effectiveness of semantic community search model.

## 5 CONCLUSION

In this paper, we present HeteroCS, a system for heterogeneous semantic community search. The system provides the functionalities of query semantic discovery, semantic-aware community search and heterogeneous community evaluation. The query semantics can be automatically discovered from the neighboring subgraph centered at the query, and represented by the frequent motifs. A community-aware GCN model is trained to generate the nodes embeddings to support the search of heterogeneous communities with different semantics. The communities are comprehensively evaluated by their query relevance, semantic consistency, and structure cohesiveness, and returned to user in a ranked order to facilitate further exploration.

## ACKNOWLEDGMENTS

This work is partially supported by the Primary Research and Development Plan of Zhejiang Province, China (2021C01164), and the advanced computing resources provided by the Supercomputing Center of Hangzhou City University.

## REFERENCES

[1] Lu Chen, Chengfei Liu, Kewen Liao, Jianxin Li, and Rui Zhou. 2019. Contextual community search over large social networks. In *2019 IEEE 35th International Conference on Data Engineering (ICDE)*. IEEE, 88–99.

[2] Wanyun Cui, Yanghua Xiao, Haixun Wang, and Wei Wang. 2014. Local search of communities in large graphs. In *Proceedings of the 2014 ACM SIGMOD international conference on Management of data*, 991–1002.

[3] Mohammed Elseidy, Ehab Abdelhamid, Spiros Skiadopoulos, and Panos Kalnis. 2014. Grami: frequent subgraph and pattern mining in a single large graph. *Proceedings of the VLDB Endowment*, 7, 7, 517–528.

[4] Yixiang Fang, Xin Huang, Lu Qin, Ying Zhang, Wenjie Zhang, Reynold Cheng, and Xuemin Lin. 2020. A survey of community search over big graphs. *The VLDB Journal*, 29, 353–392.

[5] Yixiang Fang, Yixing Yang, Wenjie Zhang, Xuemin Lin, and Xin Cao. 2020. Effective and efficient community search over large heterogeneous information networks. *Proceedings of the VLDB Endowment*, 13, 6, 854–867.

[6] Yuan Fang, Wenqing Lin, Vincent W Zheng, Min Wu, Jiaqi Shi, Kevin Chen-Chuan Chang, and Xiao-Li Li. 2019. Metagraph-based learning on heterogeneous graphs. *IEEE Transactions on Knowledge and Data Engineering*, 33, 1, 154–168.

[7] Xinyu Fu, Jiani Zhang, Ziqiao Meng, and Irwin King. 2020. Magnn: metapath aggregated graph neural network for heterogeneous graph embedding. In *Proceedings of The Web Conference 2020*, 2331–2341.

[8] Jun Gao, Jiazun Chen, Zhao Li, and Ji Zhang. 2021. Ics-gnn: lightweight interactive community search via graph neural network. *Proceedings of the VLDB Endowment*, 14, 6, 1006–1018.

[9] Xin Huang, Hong Cheng, Lu Qin, Wentao Tian, and Jeffrey Xu Yu. 2014. Querying k-truss community in large and dynamic graphs. In *Proceedings of the 2014 ACM SIGMOD international conference on Management of data*, 1311–1322.

[10] Xin Huang, Laks VS Lakshmanan, and Jianliang Xu. 2019. Community search over big graphs. *Synthesis Lectures on Data Management*, 14, 6, 1–206.

[11] Xun Jian, Yue Wang, and Lei Chen. 2020. Effective and efficient relational community detection and search in large dynamic heterogeneous information networks. *Proceedings of the VLDB Endowment*, 13, 10, 1723–1736.

[12] Xunqiang Jiang, Yuanfu Lu, Yuan Fang, and Chuan Shi. 2021. Contrastive pretraining of gnns on heterogeneous graphs. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, 803–812.

[13] Yangqin Jiang, Yixiang Fang, Chenhao Ma, Xin Cao, and Chunshan Li. 2022. Effective community search over large star-schema heterogeneous information networks. *Proceedings of the VLDB Endowment*, 15, 11, 2307–2320.

[14] Yuli Jiang, Yu Rong, Hong Cheng, Xin Huang, Kangfei Zhao, and Junzhou Huang. 2021. Qd-gcn: query-driven graph convolutional networks for attributed community search. *arXiv preprint arXiv:2104.03583*.

[15] Zemin Liu, Vincent W Zheng, Zhou Zhao, Hongxia Yang, Kevin Chen-Chuan Chang, Minghui Wu, and Jing Ying. 2018. Subgraph-augmented path embedding for semantic user search on heterogeneous social network. In *Proceedings of the 2018 World Wide Web Conference*, 1613–1622.

[16] Qiheng Mao, Zemin Liu, Chenghao Liu, and Jianling Sun. 2023. Hinormer: representation learning on heterogeneous information networks with graph transformer. *arXiv preprint arXiv:2302.11329*.

[17] Mauro Sozio and Aristides Gionis. 2010. The community-search problem and how to plan a successful cocktail party. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, 939–948.

[18] Xiao Wang, Houye Ji, Chuan Shi, Bai Wang, Yanfang Ye, Peng Cui, and Philip S Yu. 2019. Heterogeneous graph attention network. In *The world wide web conference*, 2022–2032.

[19] Xiao Wang, Nian Liu, Hui Han, and Chuan Shi. 2021. Self-supervised heterogeneous graph neural network with co-contrastive learning. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, 1726–1736.

[20] Carl Yang, Yuxin Xiao, Yu Zhang, Yizhou Sun, and Jiawei Han. 2020. Heterogeneous network representation learning: a unified framework with survey and benchmark. *IEEE Transactions on Knowledge and Data Engineering*, 34, 10, 4854–4873.

[21] Daokun Zhang, Jie Yin, Xingquan Zhu, and Chengqi Zhang. 2018. Metagraph2vec: complex semantic path augmented heterogeneous network embedding. In *Advances in Knowledge Discovery and Data Mining: 22nd Pacific-Asia Conference, PAKDD 2018, Melbourne, VIC, Australia, June 3-6, 2018, Proceedings, Part II 22*. Springer, 196–208.

[22] Wentao Zhang, Yuan Fang, Zemin Liu, Min Wu, and Xinming Zhang. 2020. Mg2vec: learning relationship-preserving heterogeneous graph representations via metagraph embedding. *IEEE Transactions on Knowledge and Data Engineering*, 34, 3, 1317–1329.

[23] Fanwei Zhu, Yuan Fang, Kevin Chen-Chuan Chang, and Jing Ying. 2013. Incremental and accuracy-aware personalized pagerank through scheduled approximation.