Efficient Cross-layer Community Search in Large Multilayer Graphs

Longxu Sun, Xin Huang, Zheng Wu, Jianliang Xu

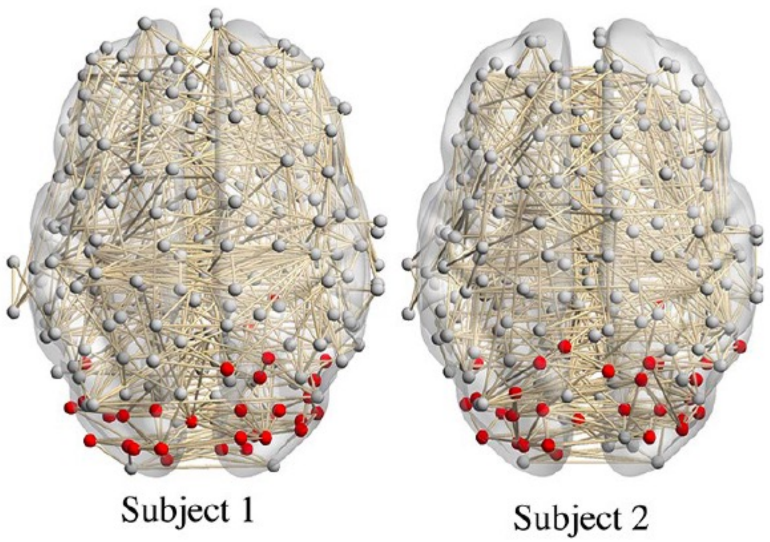
Hong Kong Baptist University

Roadmap

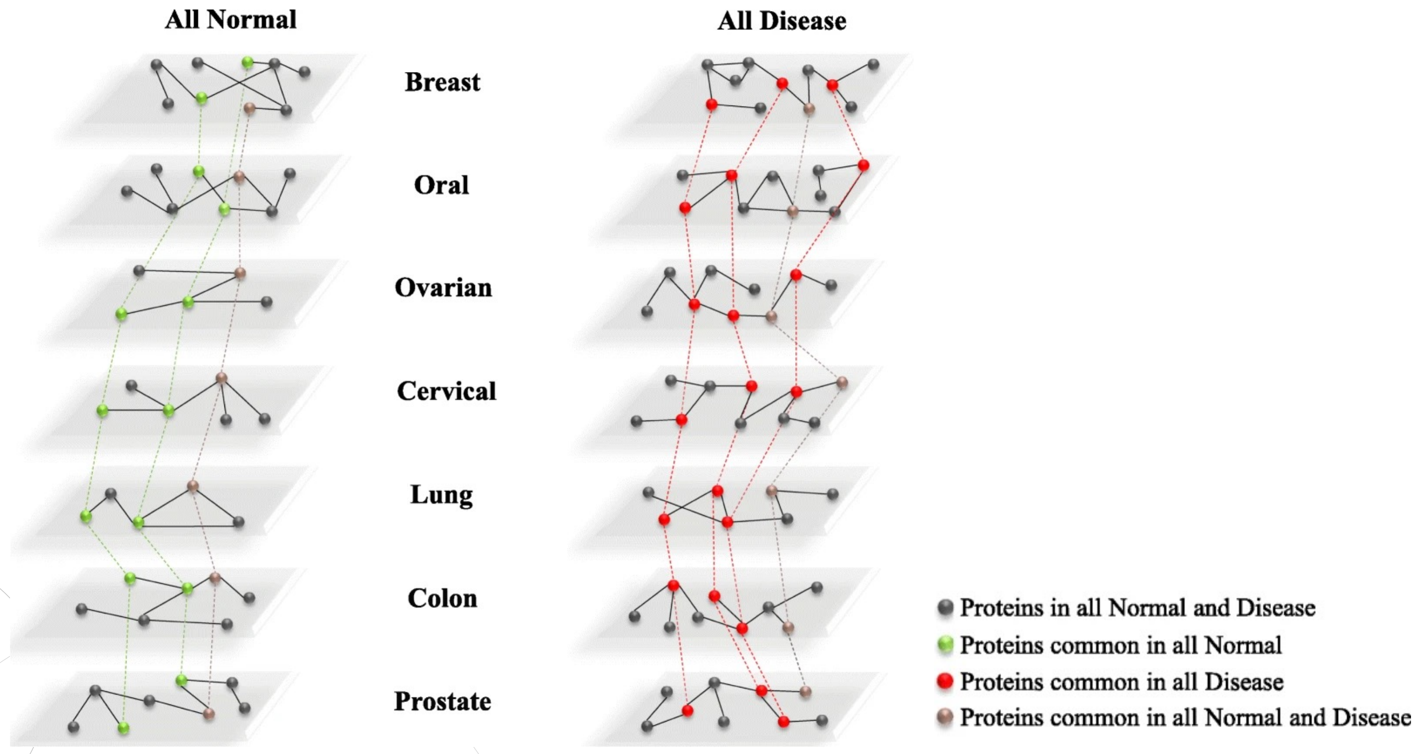
- Background
- Related Work
- Preliminaries
- Proposed Methodology
- Experiments
- Conclusions

Follow My Slides

Background: Multilayer Graphs

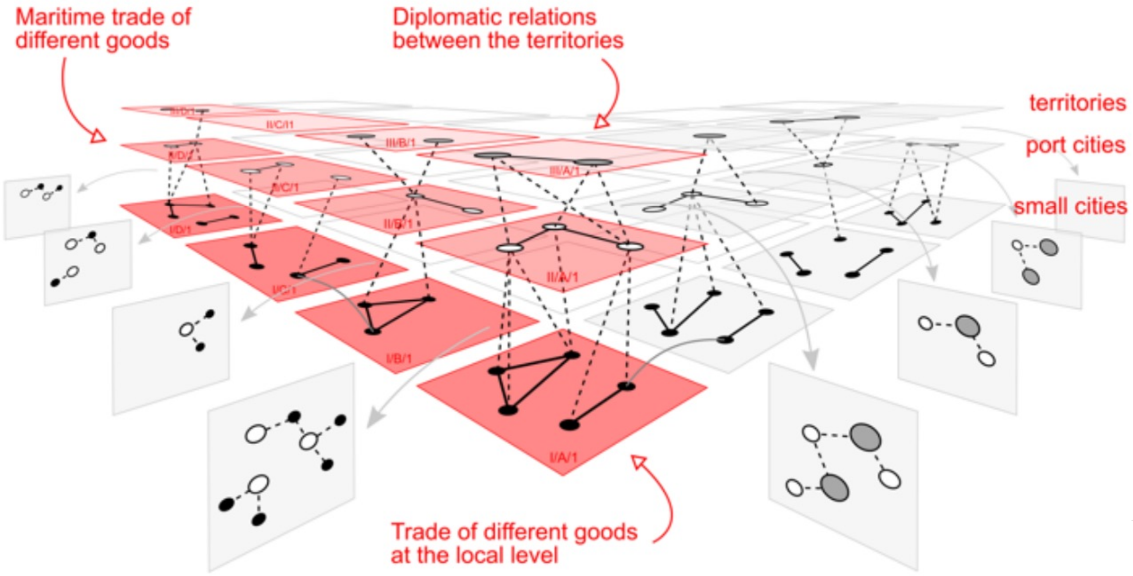


Brain Network
[Luo et al. KDD 2020]

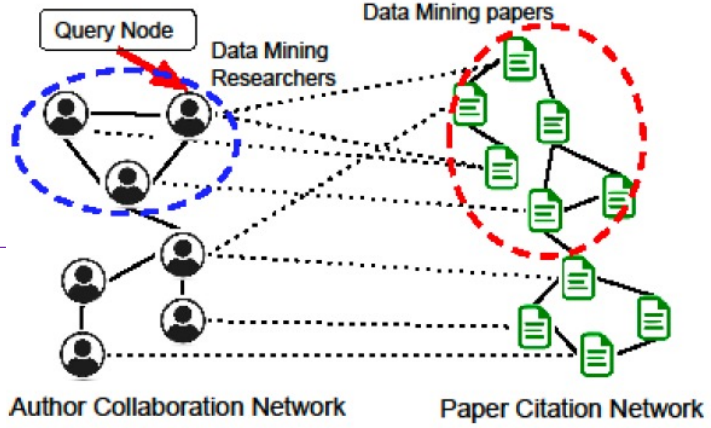


Cancer Cells' Network
[Rai et al. Sci Rep. 2017]

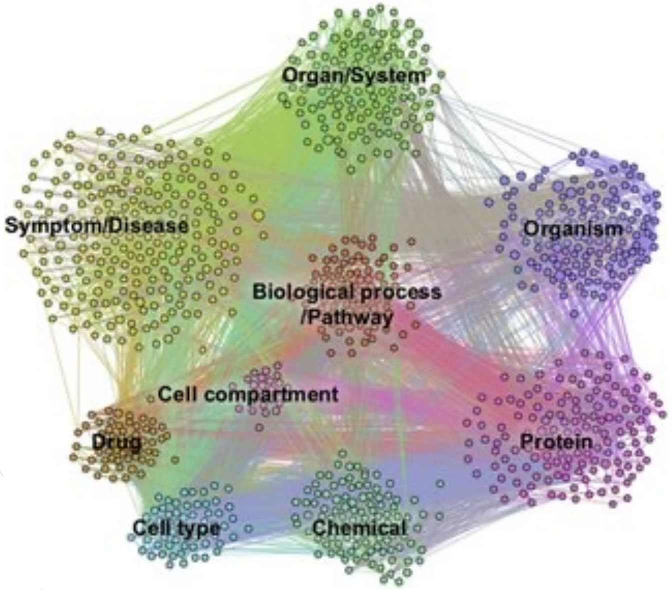
Background: Multilayer Graphs



Goods Circulation Network
[Orengo & Livarda 2016]



Citation Network
[Luo et al. KDD 2020]



Knowledge Network
[Emmanuelle et al. 2021]

Background: Community Search

Community Search: find densely connected communities

– query-dependent & highly-personalized
– support for different kinds of graphs

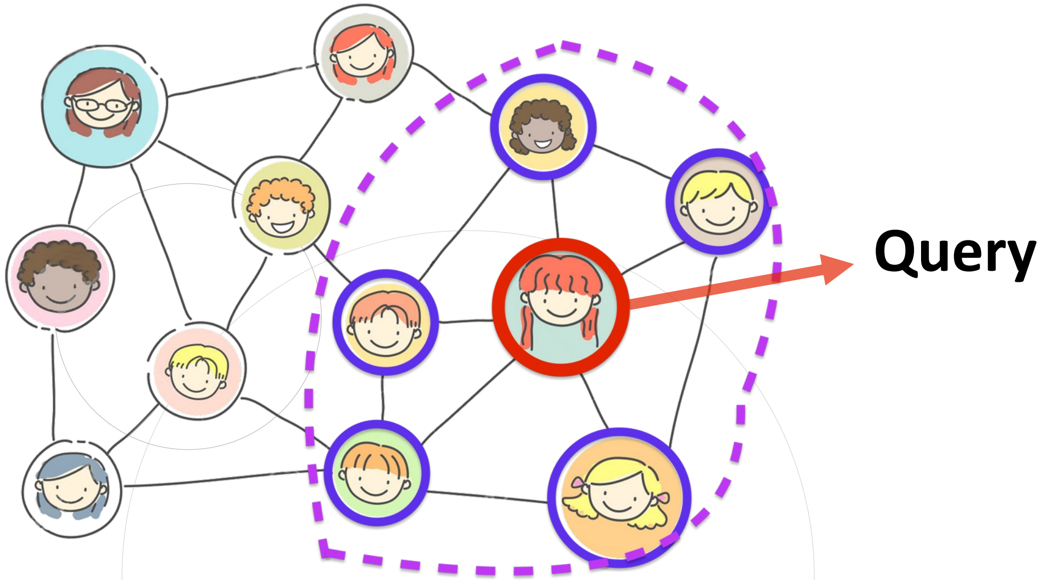
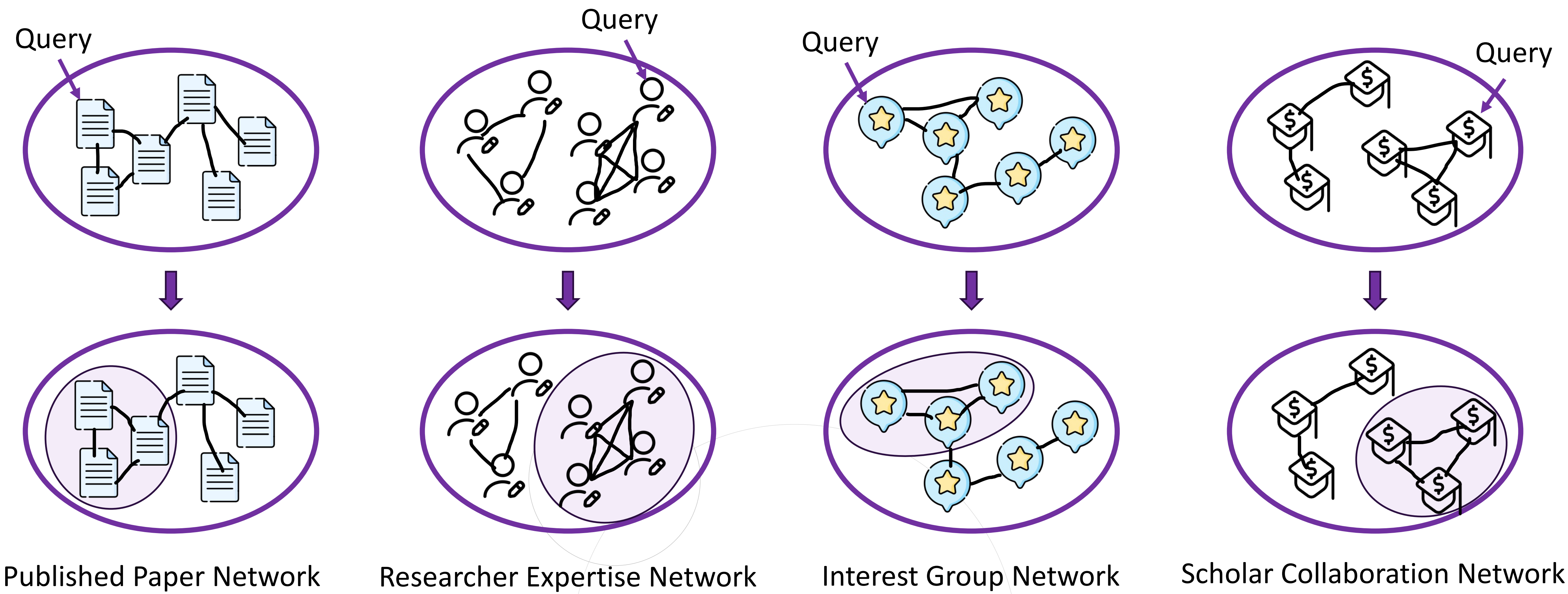


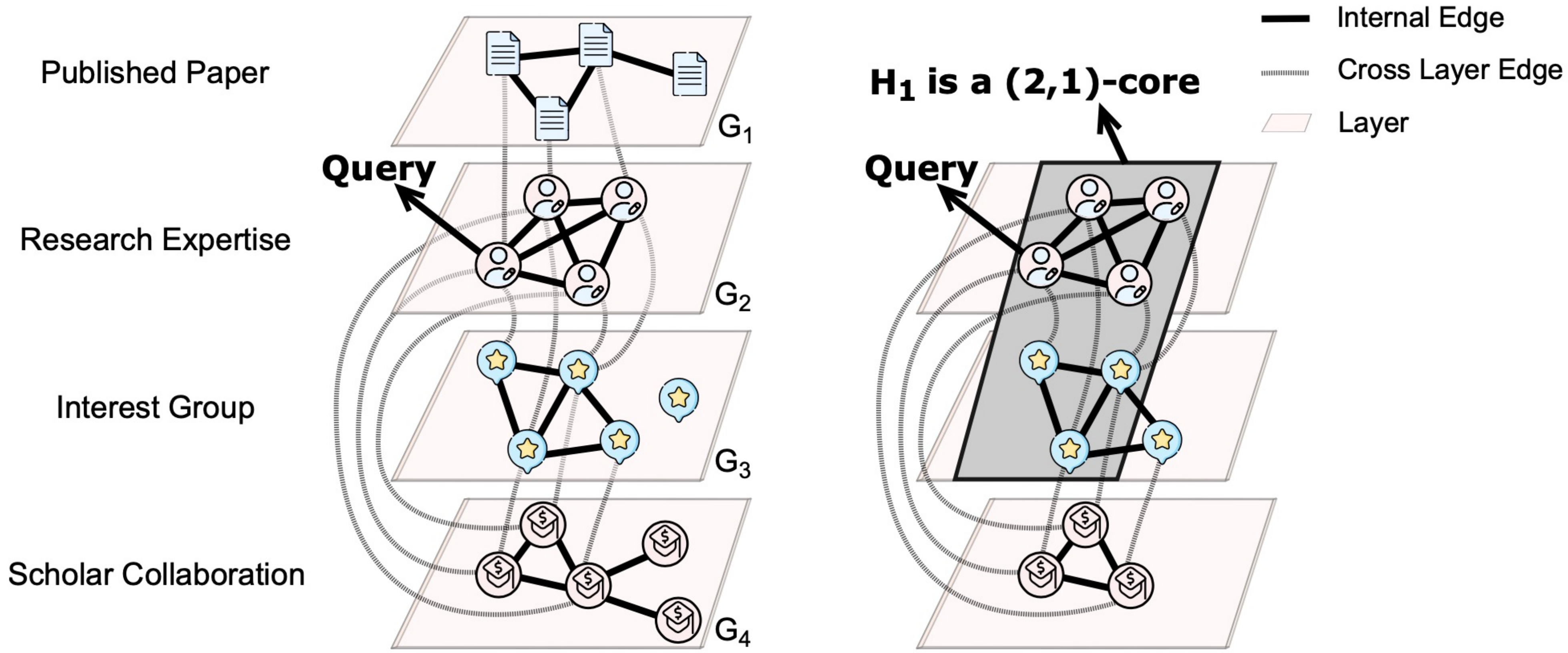
Figure Source: S. Malhotra, “Deep Dive Into Graph Traversals,” *medium.com*, 2017. [Online].

Background: Community Search



Community search in single layer graphs

Background: Community Search in MLGs

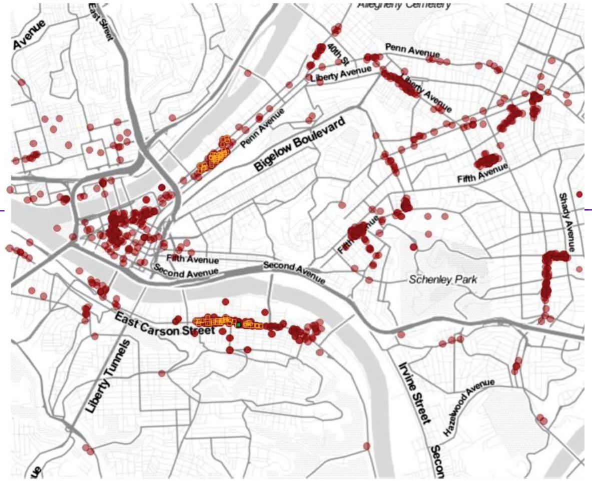


(a) Multilayer Collaboration Network MG (b) Multilayer Community H

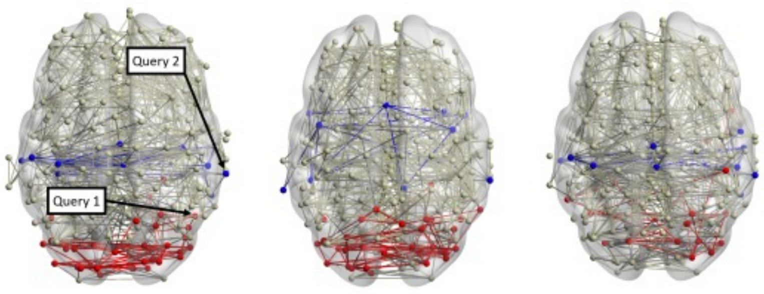
Community search in multilayer graphs

Background: Applications

- Event planning
- Social marketing
- Recommendation
- Biological data analysis

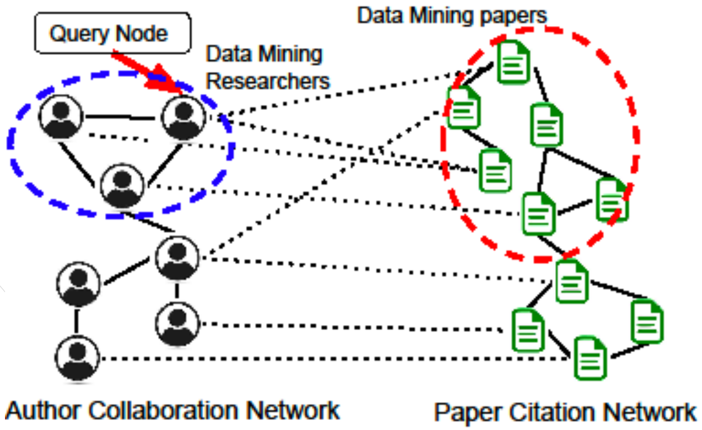


Location-based Social Marketing



(a) Subject 1 (b) Subject 2 (c) Subject 3

Identify Functional Systems



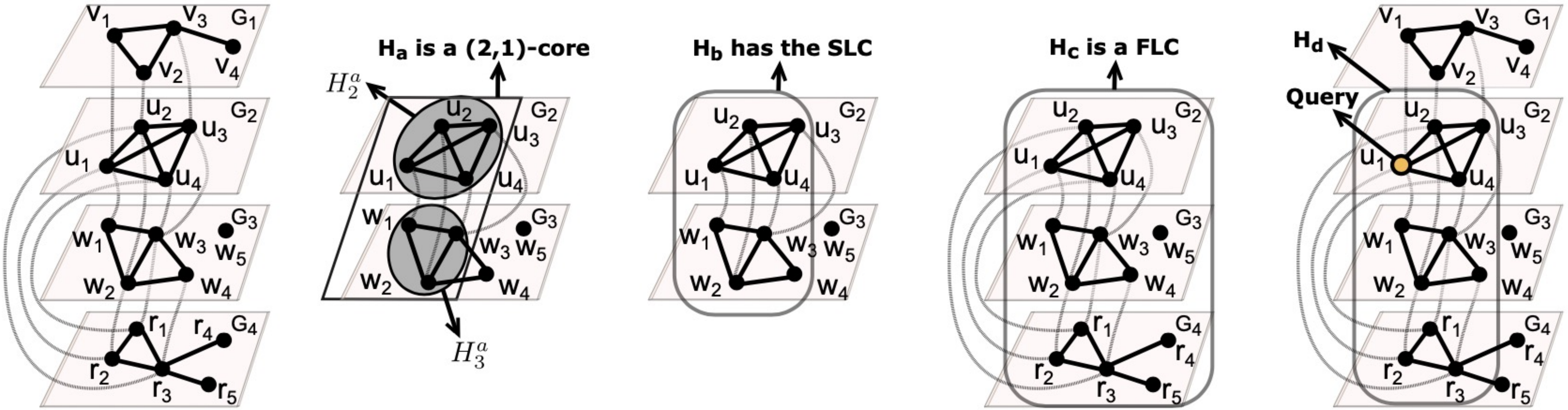
Event Planning

Related Works

1) failure to identify informative communities with the most layers when a multilayer graph is associated with a large number of layers;
2) failure to distinguish the degree of connections in internal layers and cross-layers.

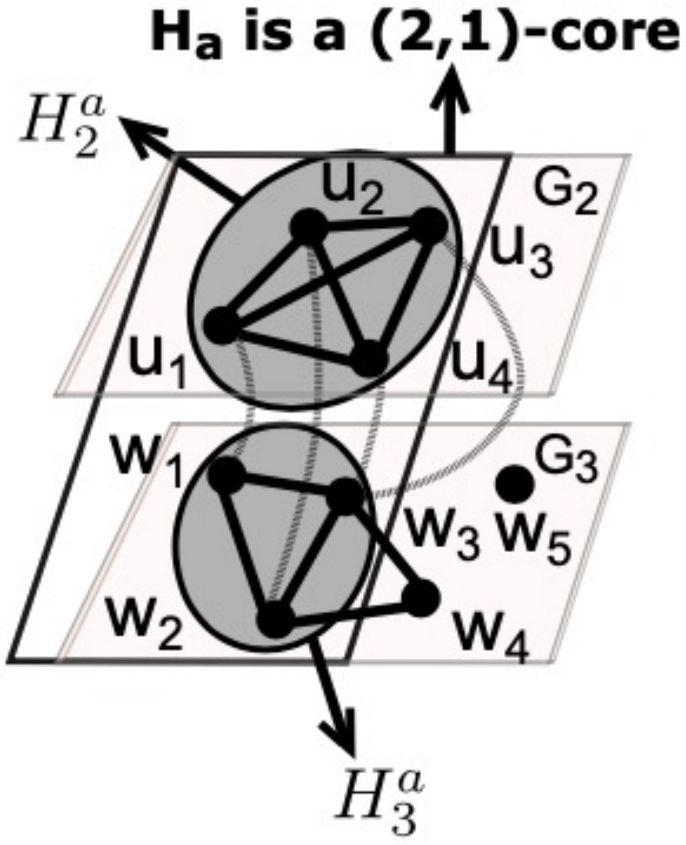
Papers	**Graph Types**			**Main Ideas**	
	Heterogeneous Information Networks	Multiplex Graphs	Multilayer Graphs	Dense Subgraph Model	Random Walk
Fang et al. PVLDB 2020	✓			✓	
Jiang et al. PVLDB 2022	✓			✓	
Zhou et al. PVLDB 2023	✓			✓	
Behrouz et al. PVLDB 2022		✓		✓	
Luo et al. KDD 2020	✓	✓	✓		✓
Ours	✓	✓	✓	✓	

Preliminary: Overview



(a) Multilayer Graph MG (b) (k, d)-core H_a (c) Strong Cross-layer Connectivity (d) Full-connected Multilayer Community (e) Multilayer Community Search

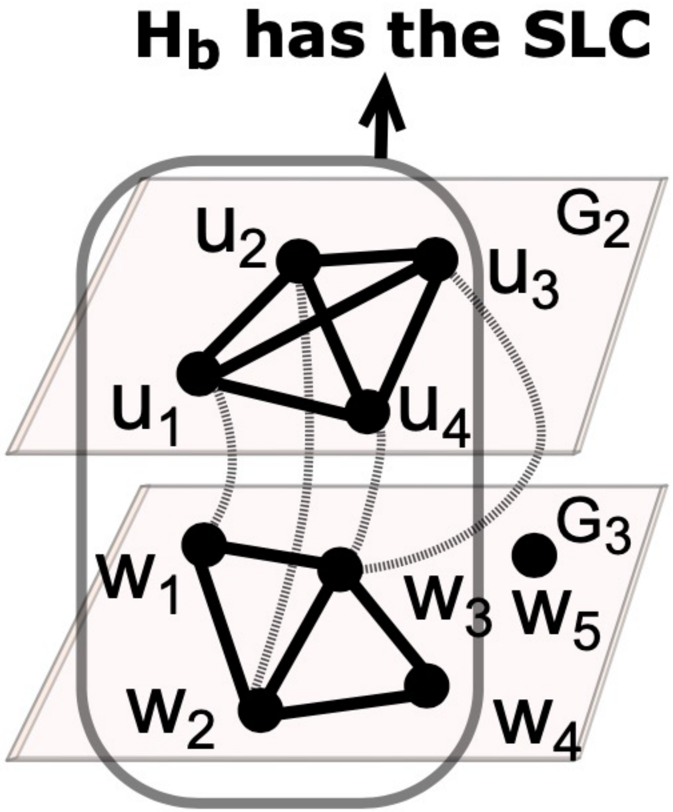
Preliminary: (k, d)-core



Definition 1 ((k, d)-core). Given a multilayer graph MG and two parameters $k, d \in \mathbb{Z}^{\geq 0}$, a connected two-layer subgraph $H(H_i, H_j, E_{ij}^H) \subseteq$ MG located at layers G_i and G_j is (k, d)-core *if and only if* H admits the following conditions:

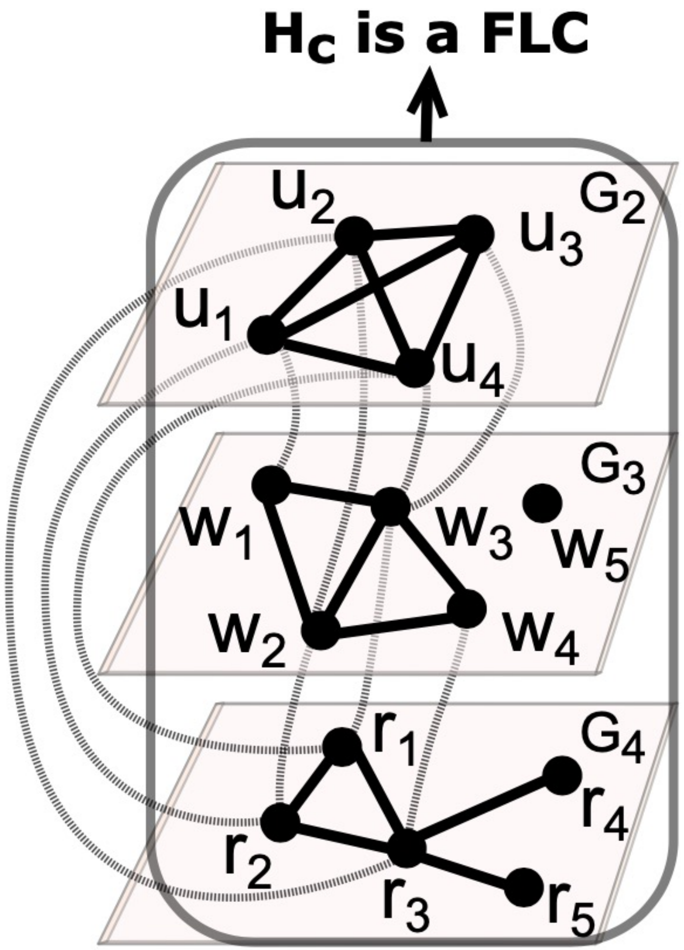
1) $\forall v \in V_i(H)$, the intra-degree at layer G_i: $\deg_{H_i}(v) \geq k$;
2) $\forall v \in V_j(H)$, the intra-degree at layer G_j: $\deg_{H_j}(v) \geq k$;
3) $\forall v \in V_i(H) \cup V_j(H)$, the inter-degree: $\deg_{E_{ij}^H}(v) \geq d$.

Preliminary: Strong Cross-layer Connectivity



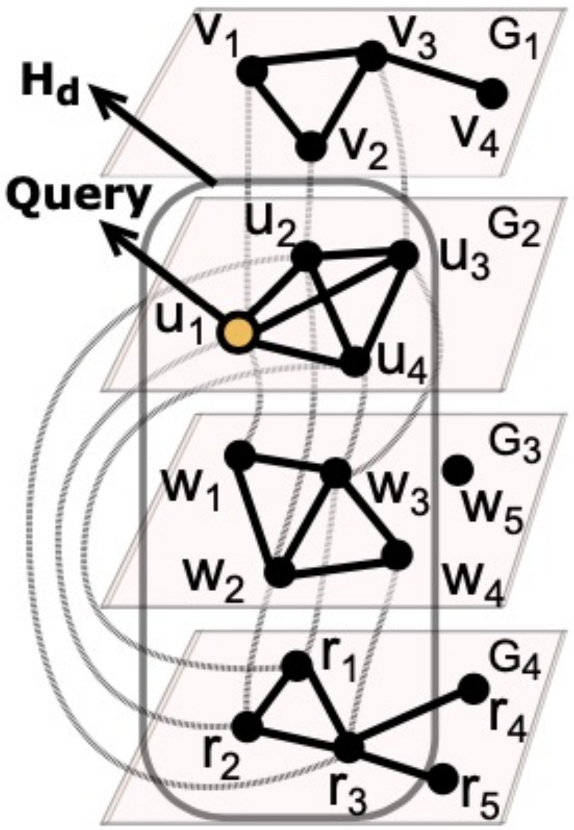
Definition 2 (Strong Cross-layer Connectivity)**.** Given a multilayer graph H, we say that H has the *strong cross-layer connectivity* between two layers G_i and G_j *if and only if* there exists a non-empty two-layer subgraph of (k, d)-core $H' \subseteq H$ at layers G_i, G_j, denoted as $H_i \xleftrightarrow{H} H_j$.

Preliminary: Fully-connected Multilayer Connectivity



Definition 3 (Fully-connected Multilayer Community). Given a multilayer subgraph $H \subseteq \mathrm{MG}$ and two numbers k, d, we say that H is a full-layer connected multilayer community *if and only if* for every pair of layers $i, j \in \mathcal{L}(H)$, there exists a strong cross-layer connectivity between G_i and G_j, such that, $\forall i, j \in \mathcal{L}(H)$, $H_i \xleftrightarrow{H} H_j$.

Preliminary: Multilayer Community Search Problem



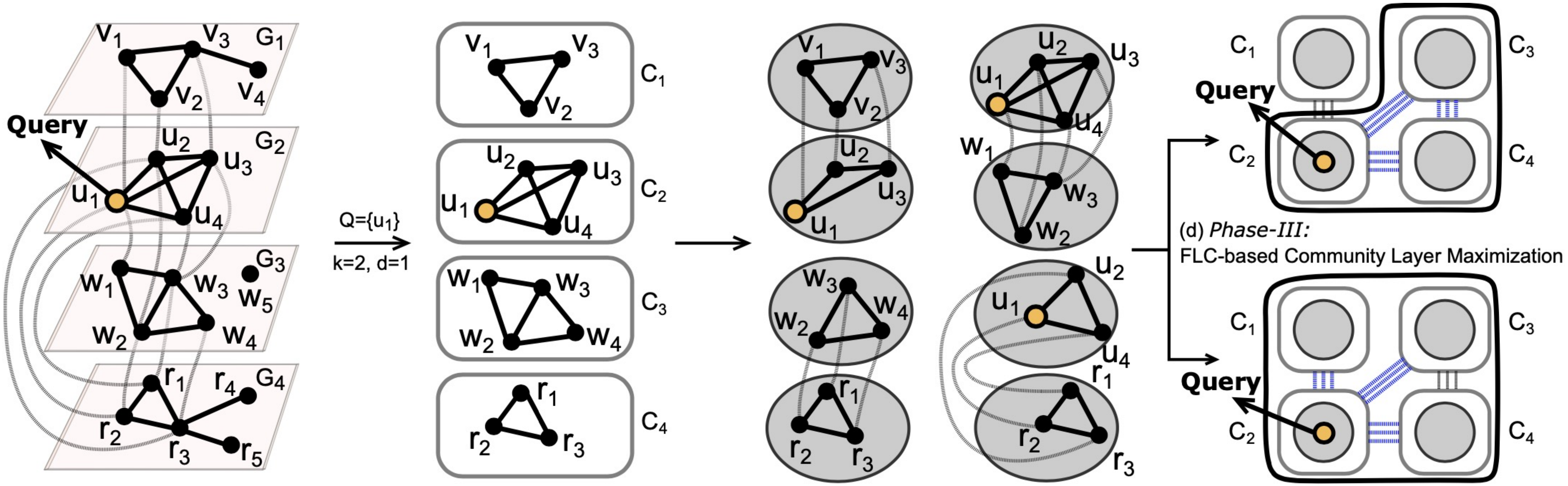
MCS-problem is NP-Hard

Problem 1 (MCS-problem). *Given a multilayer graph* MG$(V_M, E_M, \mathcal{L})$, *a set of query vertices* $Q \subseteq V_M$, *two parameters* $k, d \in \mathbb{Z}^{\geq 0}$, *the problem of cross-layer community search in* MG (MCS-problem) *is to find a connected community* $H \subseteq$ MG *satisfying the following four constraints:*

1) ***Query-dependent personalization:*** $Q \subseteq V(H)$;
2) ***Core-dense internal layers:*** $\forall i \in \mathcal{L}(H)$, H_i *is a connected k-core;*
3) ***Fully-connected cross-layers:*** $\forall i, j \in \mathcal{L}(H)$, *two layers H_i and H_j are connected via a (k, d)-core in H;*
4) ***Cross-layer maximization:*** $|\mathcal{L}(H)|$ *is maximized.*

Proposed Methodologies: Framework

The additional phase is for relaxed MCS problem which starting from Q to search path-connected multilayer community H.

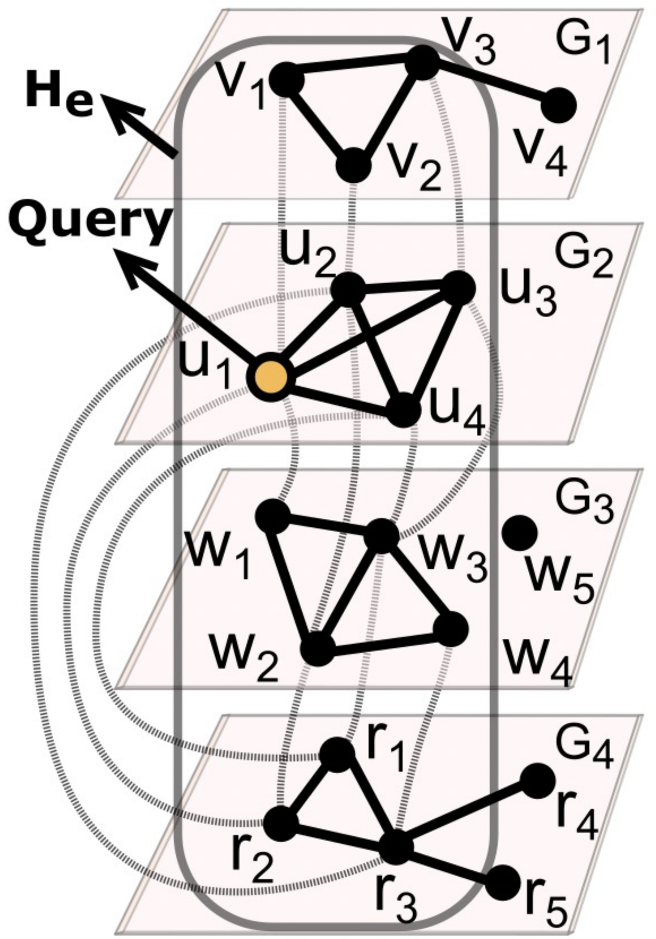


(a) Multilayer Graph MG (b) *Phase-I:* K-Core Component Extraction (c) *Phase-II:* Cross-layer (k, d)-Core Verification (e) Path-layer based Community Search

The key idea of MCS framework consists of three phases:

I. extracting k-core components at all layers;
II. identifying the strong cross-layer connectivity for any two k-core components at different layers;
III. starting from Q to search fully-connected multilayer community H.

Proposed Methodologies: Relaxed pMCS-problem



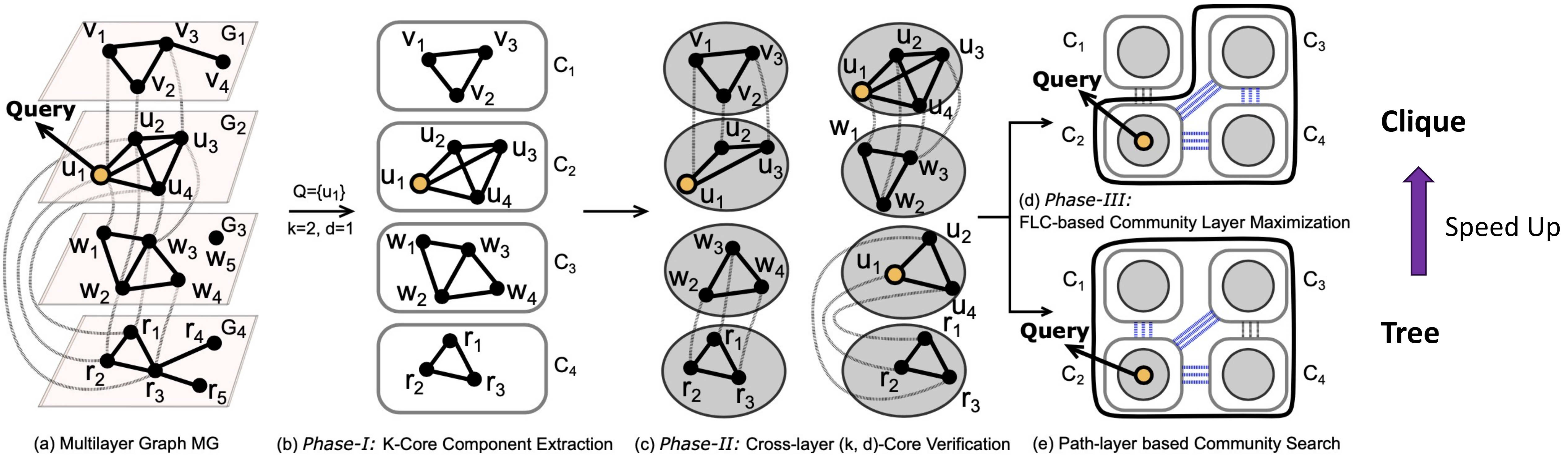
Definition 4 (Path-layer Connectivity)**.** For a given multilayer subgraph $H \subseteq$ MG, two layers H_i and H_j with $i, j \in \mathcal{L}(H)$, $i \neq j$, we say that H_i and H_j has the *path-layer connectivity*, denoted as $H_i \overset{H}{\leftrightsquigarrow} H_j$, *if and only if* there exists a path $(H_{p_1}, \ldots, H_{p_r})$ such that every pair of layers $(H_{p_x}, H_{p_{x+1}})$ where $1 \leq x < r$, is *strong cross-layer connected*, i.e., $H_{p_x} \overset{H}{\longleftrightarrow} H_{p_{x+1}}$, $p_1 = i$, and $p_r = j$.

Problem 3 (pMCS-problem)**.** *Given a multilayer graph* MG*, a set of query vertices* Q*, two parameters* k *and* d*, the problem of path-layer based multilayer community search is to find a connected subgraph* $H \subseteq$ MG *satisfying four constraints:*

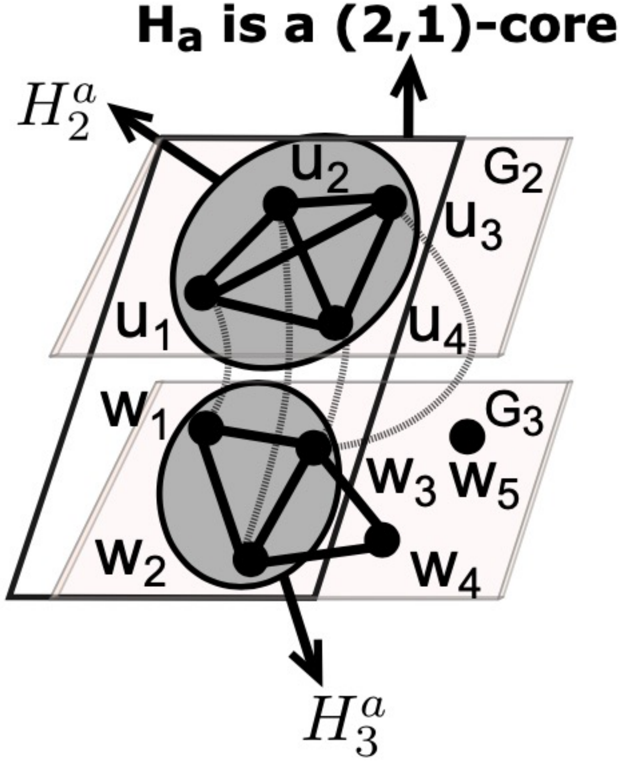
1) *H contains all query vertices Q;*
2) *$\forall i \in \mathcal{L}(H)$, H_i is a k-core;*
3) *$\forall i, j \in \mathcal{L}(H)$, the path-layer connectivity $H_i \overset{H}{\leftrightsquigarrow} H_j$ always holds;*
4) *$|\mathcal{L}(H)|$ is maximized.*

Tree Structure

Proposed Methodologies: MCS vs pMCS



Proposed Methodologies: (k, d)-core Index



Definition 5 ((k,d)-coreness)**.** Given two layers G_i and G_j, the (k,d)-coreness of a cross-layer edge (u,v) in E_{ij}, denoted as $\Phi((u,v))$, is a set of unique (k,d) pairs, such that for any pair of $(k,d) \in \Phi((u,v))$, $\Phi^k((u,v)) = d$. Moreover, for any $(k,d) \in \Phi((u,v))$, there exists no $(k',d') \in \Phi((u,v))$ such that both $k' \geq k$ and $d' \geq d$ hold.

Experiments: Datasets

Setup: All experiments were performed on a server with an Intel Xeon Gold 6330 2.0 GHz CPU and 1T RAM, running 64-bit Oracle Linux 8.8.

Datasets types:

Collaboration networks: DBLP, Citeseer

Social networks: Twitter, Friendfeed, Venetie

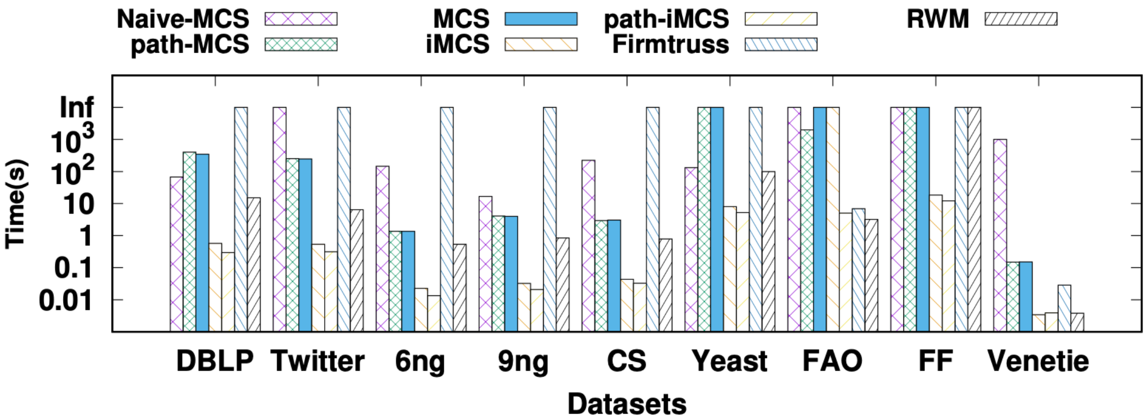
News groups: 6ng, 9ng

Protein-protein networks: Yeast

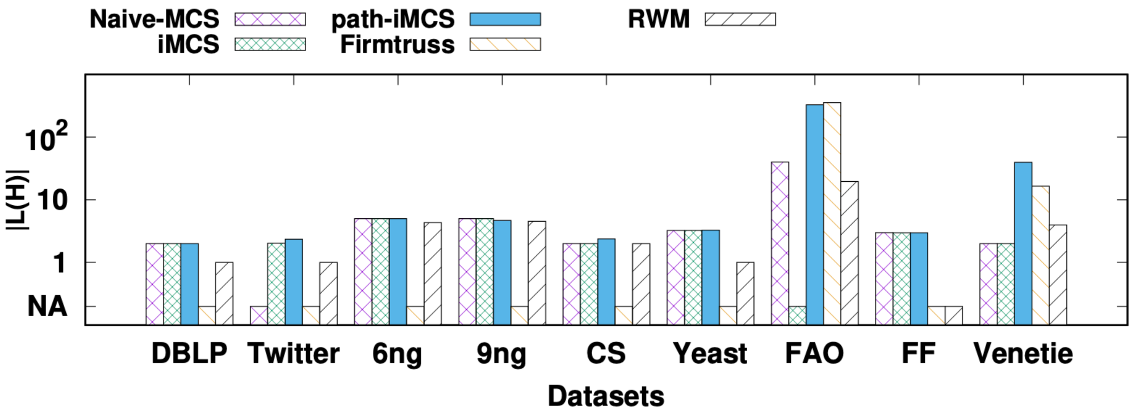
Agriculture data: FAO

Dataset	n	m	m_l	m_s	l
DBLP	41,892	661,883	280,707	381,176	2
Twitter	47,280	535,062	445,287	89,775	3
6ng	4,500	29,984	9,000	20,984	5
9ng	6,750	44,980	13,500	31,480	5
Citeseer (CS)	15,533	68,376	56,548	11,828	3
Yeast	4,458	8,500,745	8,473,997	26,748	4
FAO	214	14,456,470	318,346	14,138,124	364
FriendFeed (FF)	510,338	20,204,534	18,673,520	1,531,014	3
Venetie	206	21,310	19,955	1,355	43

Experiments: Comparison on All Algorithms



Efficiency evaluation on all chosen datasets.



Quality evaluation on all chosen datasets.

RWM: a random-walk-based approach for local multilayer community search [Luo et al. KDD 2020].

FirmTruss: a truss-based community search approach in multiplex networks [Behrouz et al. PVLDB 2022].

Naive-MCS: our baseline method for full-layered community enumerations.

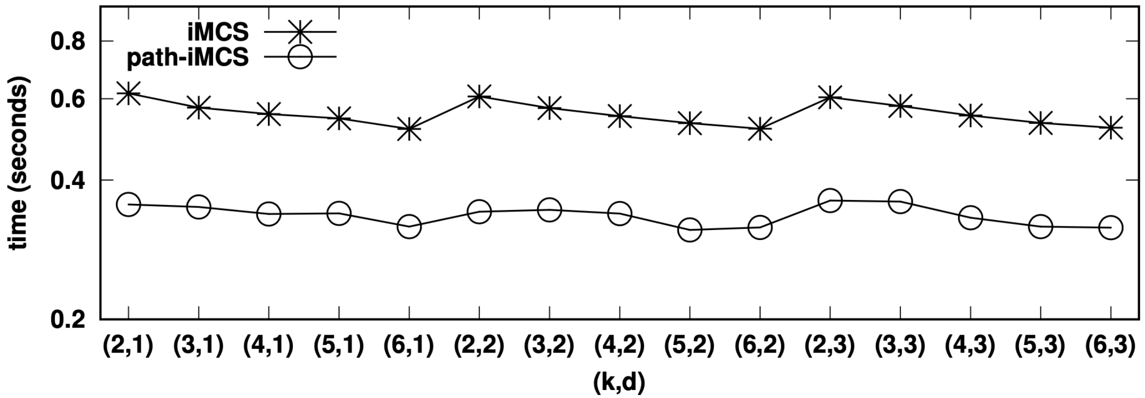
Path-MCS: the path-layer community search algorithm.

MCS: our fast approach for full-layer community search.

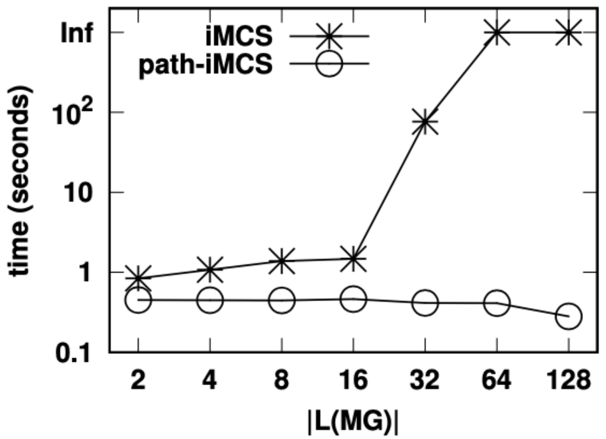
Path-iMCS: our index-based Path-MCS.

iMCS: our index-based improved approach of MCS.

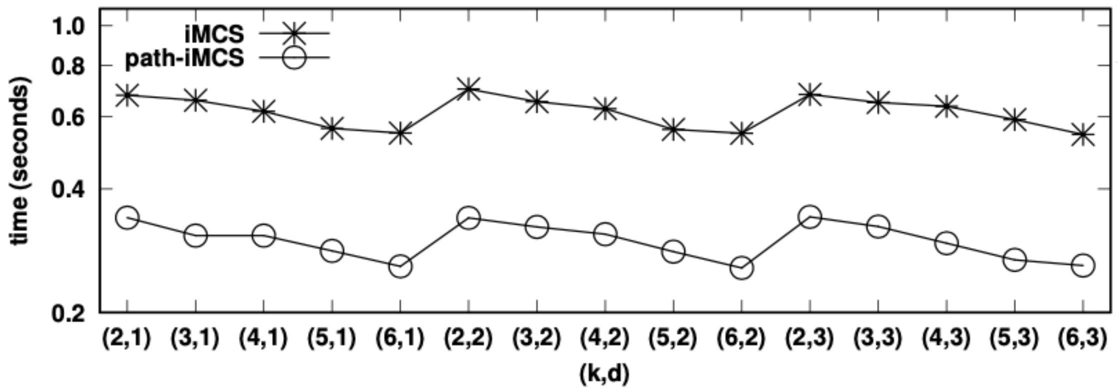
Experiments: Scalability Evaluation



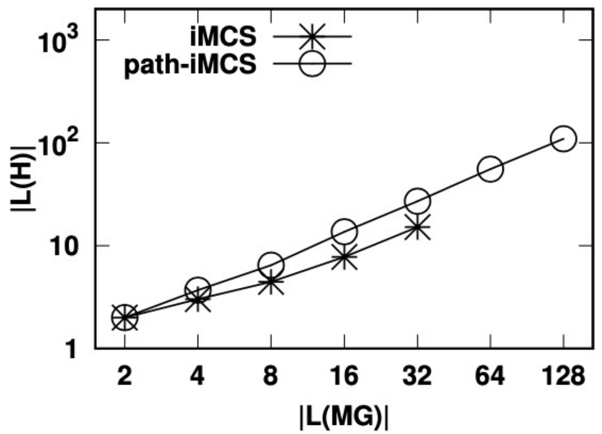
Efficiency evaluation by varying parameters (k, d) on Twitter.



Efficiency evaluation by varying |L(MG)|.



Efficiency evaluation by varying parameters (k, d) on DBLP.



Layer number evaluation by varying |L(MG)|.

Experiments: Index Quality Comparison

A COMPARISON OF INDEX SIZE AND CONSTRUCTION TIME.

	DBLP	Twitter	6ng	9ng	CS	Yeast	FAO	FF	Venetie
Graph Size (MB)	8.99	9.43	0.377	0.59	1.71	117	211	357	0.286
(k, d)-core Index Size (MB)	4.15	7.32	0.25	0.376	1.1	336	523	307	0.417
FirmTruss Index Size (MB)	NA	NA	NA	NA	NA	1380	2600	2670	2.8
(k, d)-core-indexing Time (h)	3.63	3.44	0.005	0.009	1	29.02	4.27	98.39	0.001
FirmTruss-indexing Time (h)	NA	NA	NA	NA	NA	15.54	0.71	11.55	0.008

Conclusion

- We propose a novel dense subgraph of **(k, d)-core** in multilayer graphs, strengthening the connections in internal layers and cross-layers.
- Based on (k, d)-core, we formulate our new problem of **multilayer community search (MCS)** to maximize the number of cross-layers and prove the MCS problem to be NP-hard.
- We propose an exact enumeration algorithm and a bound-and-search method, which reduces the *full-layer connectivity* to *path-layer connectivity* to accelerate efficiency.
- In addition, we design a **(k, d)-core index** to store all (k, d)-core information and propose an index-based algorithm to speed up community search.
- We conduct extensive experiments to evaluate the effectiveness and efficiency of our algorithms on **nine** real-world datasets.

THANK YOU

Q&A

LEAVE YOUR COMMENT

VIEW MY HOMEPAGE

VIEW MY SLIDES