

FedGraph for Health

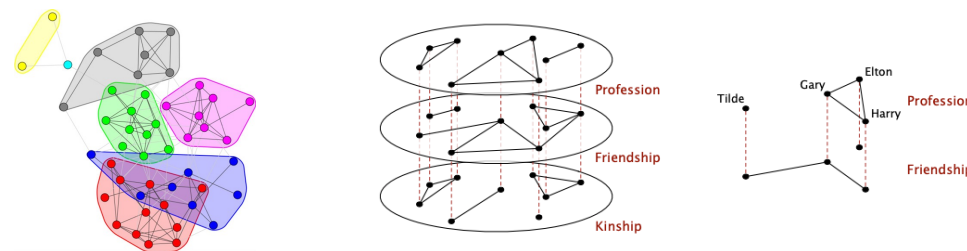
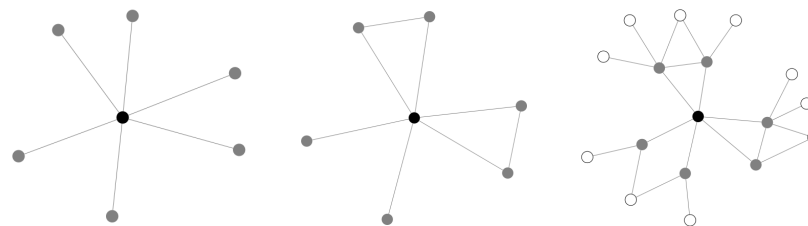
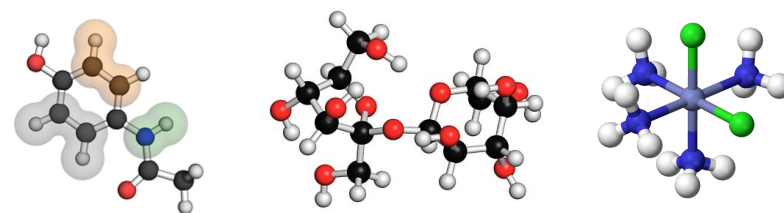
Carl Yang@Emory

FedGraph: 3 Settings

- Graph-level FL
 - Each pharmaceutical research institution holds a set of graphs
 - E.g., Drug discovery, molecular trial, protein synthesis, brain network analysis
- Subgraph-level FL
 - Each institute holds a subset of a global graph
 - E.g., Knowledge graph completion, diagnosis prediction
- Link-level FL
 - Each client holds a set of links among the same set of nodes
 - E.g., Social network analysis, recommendation

Need: jointly train a powerful and generic model

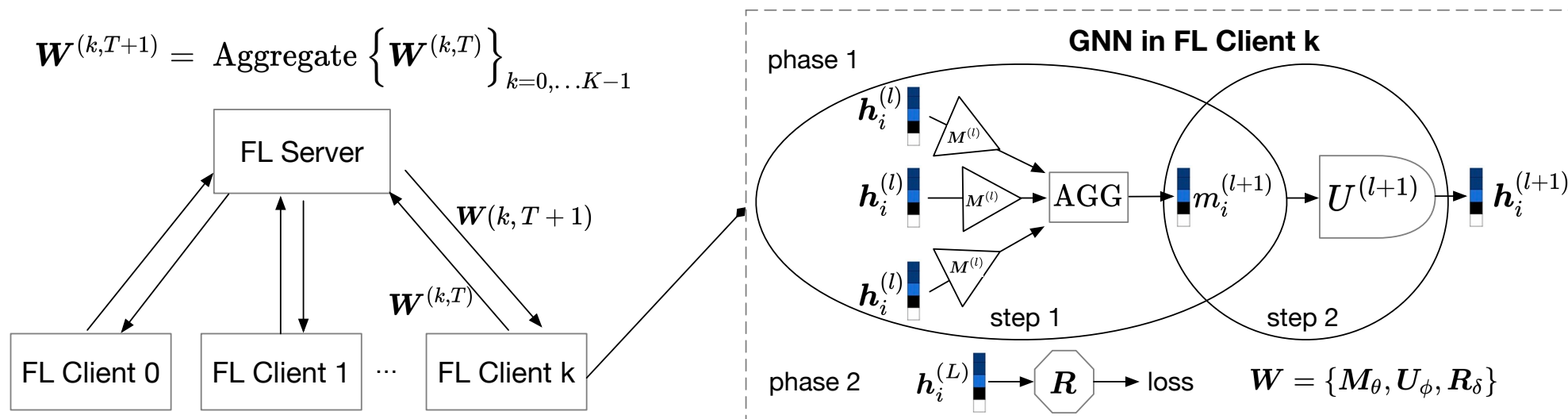
Reality: Privacy concerns, regulatory restrictions, commercial competitions



FedGraph: 3 Challenges

- Utility: Approximating centralized graph learning
 - Graph-level: non-iid structures
 - Subgraph-level: incomplete neighborhoods
 - Link-level: heterogeneous links
- Efficiency: Reducing graph computation overhead
 - Graph-level: asynchronous graph model aggregation
 - Subgraph-level: multimodal communication compression
 - Link-level: partial link model aggregation
- Privacy: Guaranteeing cross-client graph privacy
 - Interactive DP: Ensuring DP by leveraging aggregation/compression
 - Non-interactive DP: Ensuring DP by private graph generation

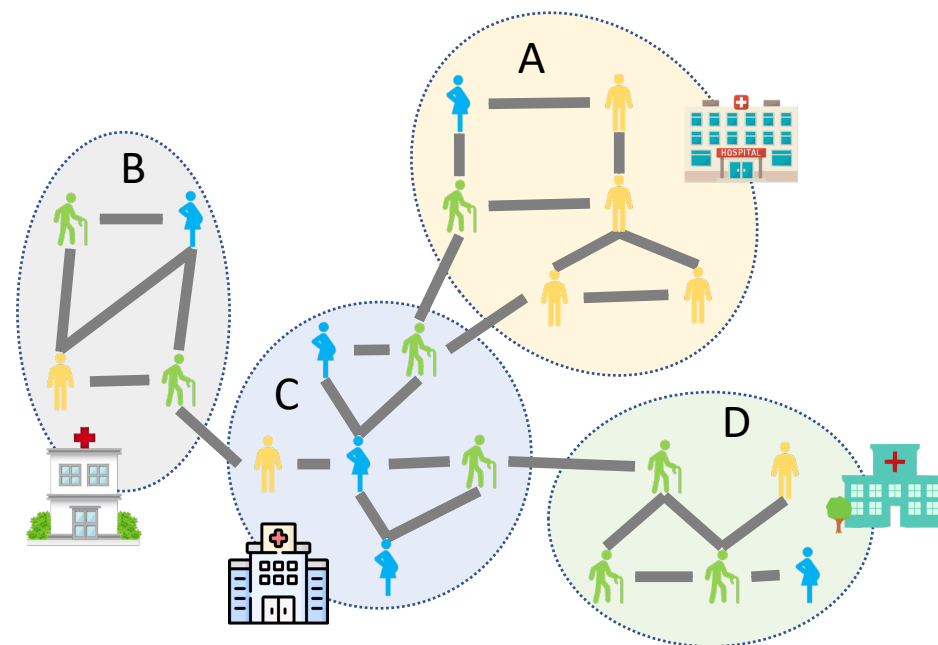
FedGraph: Benchmark (ICLR/MLSys'21)



- GNN for graph learning: message passing + readout
- Federated learning: share gradients instead of data
- FedGraphNN: popular methods, 36 open datasets, 1 distributed system

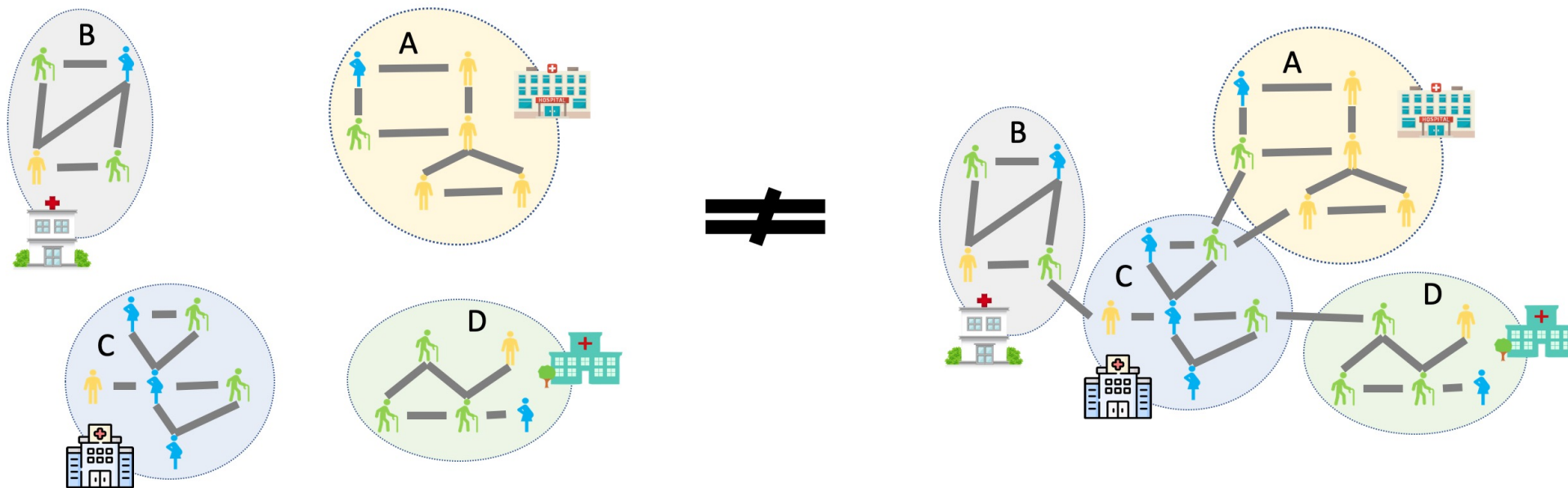
FedGraph Example: Utility of Subgraph-Level FL

- Hospitals capture local subgraphs of patient interactions
- City-level disease control needs unbiased global predictions
- Restrictions on direct data transmission/integration



Main Challenge

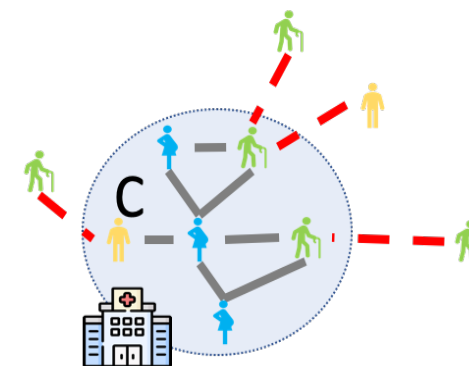
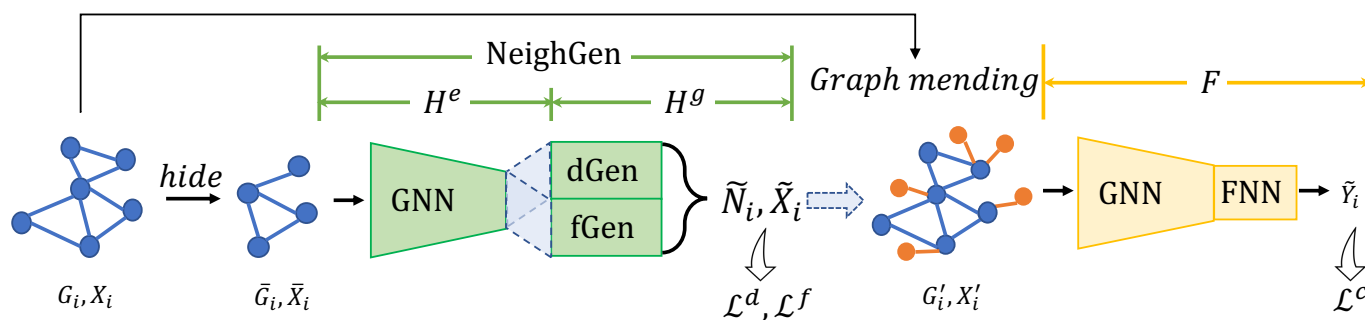
How to handle cross-subgraph missing links?



Our Solution:

FedSage + NeighGen [NeurIPS'21 Spotlight]

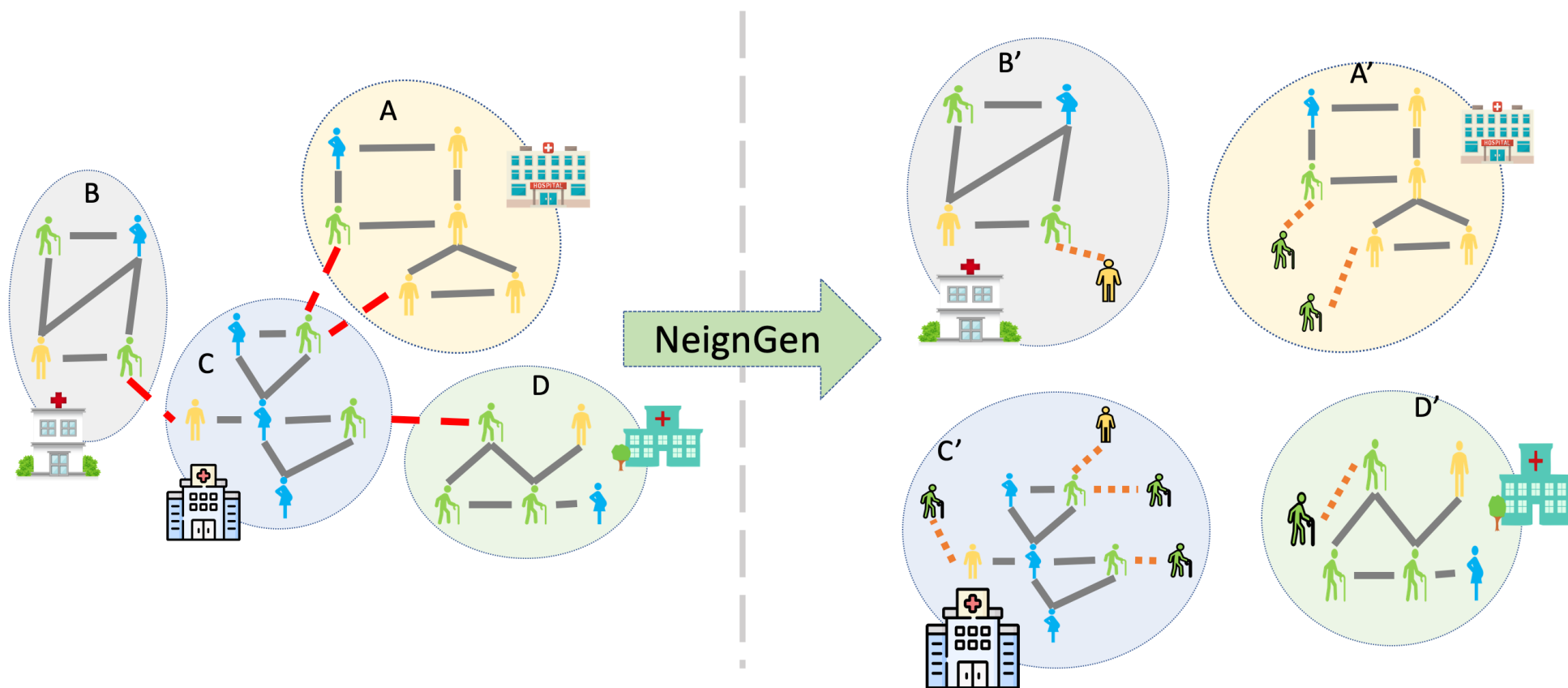
- NeighGen: Missing neighbor generator



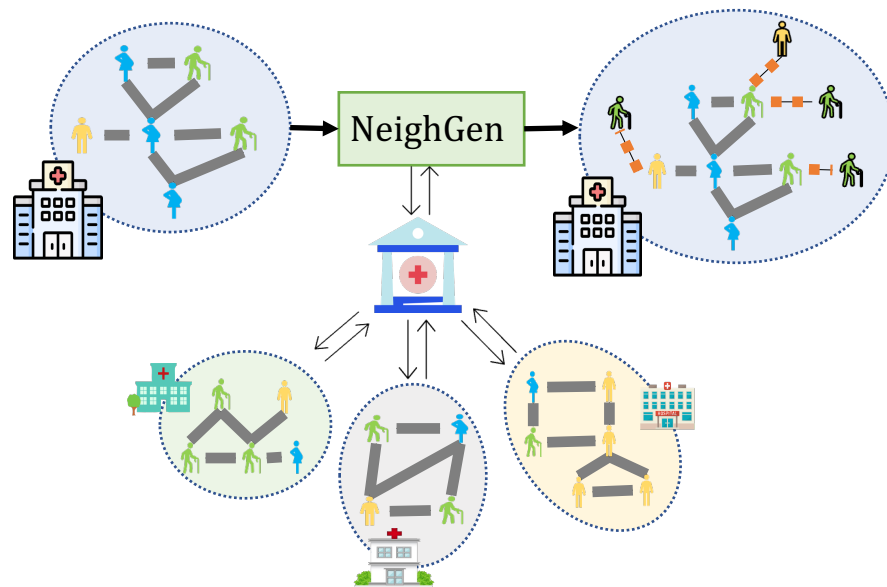
$$\tilde{N}_i = \{\tilde{n}_v | \tilde{n}_v \in \mathbb{N}, v \in V_i\}$$

$$\tilde{X}_i = \{\tilde{x}_v | \tilde{x}_v \in \mathbb{R}^{\tilde{n}_v \times x}, \tilde{n}_v \in \tilde{N}_i, v \in V_i\}$$

NeighGen: Ideal Effect



NeighGen: Federated Training



Communication overheads
& potential privacy issues...

$$\mathcal{L}_i^f = \frac{1}{|\bar{V}_i|} \sum_{v \in \bar{V}_i} \sum_{p \in [\tilde{n}_v]} \left(\min_{u \in \mathcal{N}_{G_i}(v) \cap V_i^h} (\|\tilde{x}_v^p - x_u\|_2^2) + \alpha \sum_{j \in [M]/i} \min_{u \in V_j} (\|H_i^g(z_v)^p - x_u\|_2^2) \right)$$

Experimental Results

Baselines:

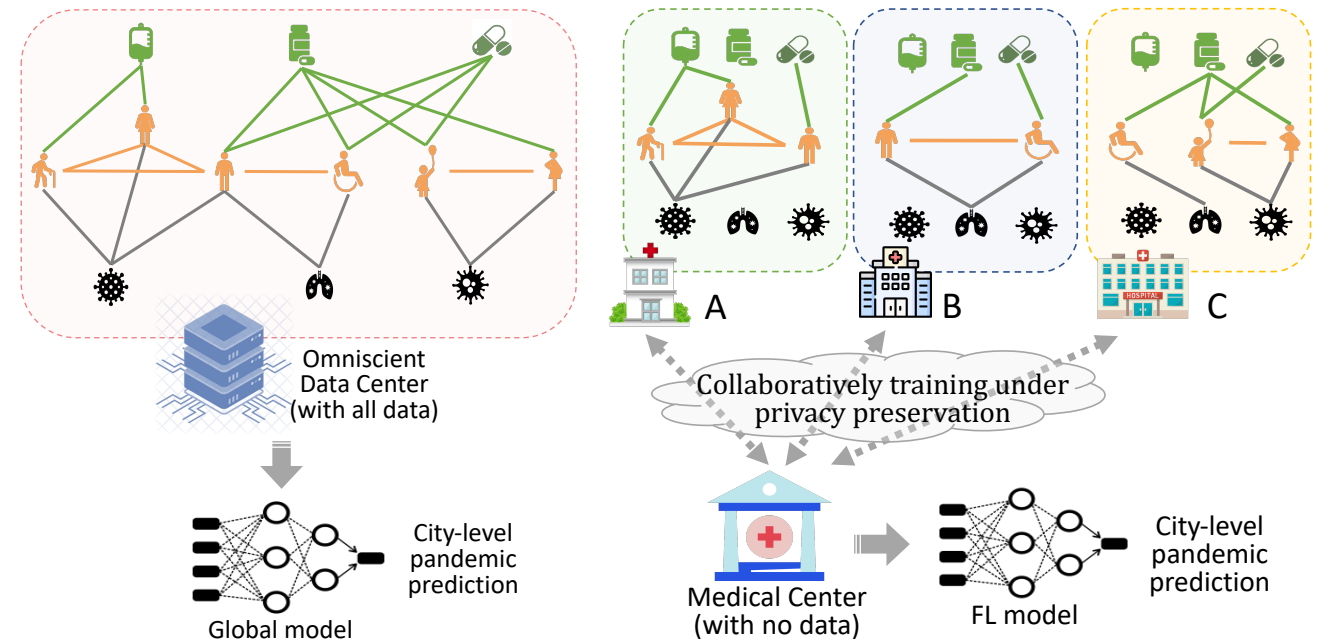
GlobSage: the GraphSage model trained on the original global graph without missing links (upper bound)

LocSage: one GraphSage model trained solely on each subgraph

Model	Cora			Citesser		
	M=3	M=5	M=10	M=3	M=5	M=10
LocSage	0.5762 (± 0.0302)	0.4431 (± 0.0847)	0.2798 (± 0.0080)	0.6789 (± 0.054)	0.5612 (± 0.086)	0.4240 (± 0.0859)
LocSage+	0.5644 (± 0.0219)	0.4533 (± 0.047)	0.2851 (± 0.0080)	0.6848 (± 0.0517)	0.5676 (± 0.0714)	0.4323 (± 0.0715)
FedSage	0.8656 (± 0.0043)	0.8645 (± 0.0050)	0.8626 (± 0.0103)	0.7241 (± 0.0022)	0.7226 (± 0.0066)	0.7158 (± 0.0053)
FedSage+	0.8686 (± 0.0054)	0.8648 (± 0.0051)	0.8632 (± 0.0034)	0.7454 (± 0.0038)	0.7440 (± 0.0025)	0.7392 (± 0.0041)
GlobSage	0.8701 (± 0.0042)			0.7561 (± 0.0031)		
Model	PubMed			MSAcademic		
	M=3	M=5	M=10	M=3	M=5	M=10
LocSage	0.8447 (± 0.0047)	0.8039 (± 0.0337)	0.7148 (± 0.0951)	0.8188 (± 0.0331)	0.7426 (± 0.0790)	0.5918 (± 0.1005)
LocSage+	0.8481 (± 0.0041)	0.8046 (± 0.0318)	0.7039 (± 0.0925)	0.8393 (± 0.0330)	0.7480 (± 0.0810)	0.5927 (± 0.1094)
FedSage	0.8708 (± 0.0014)	0.8696 (± 0.0035)	0.8692 (± 0.0010)	0.9327 (± 0.0005)	0.9391 (± 0.0007)	0.9262 (± 0.0009)
FedSage+	0.8775 (± 0.0012)	0.8755 (± 0.0047)	0.8749 (± 0.0013)	0.9359 (± 0.0005)	0.9414 (± 0.0006)	0.9314 (± 0.0009)
GlobSage	0.8776 (± 0.0011)			0.9681 (± 0.0006)		

Extension to Heterogeneous Graphs

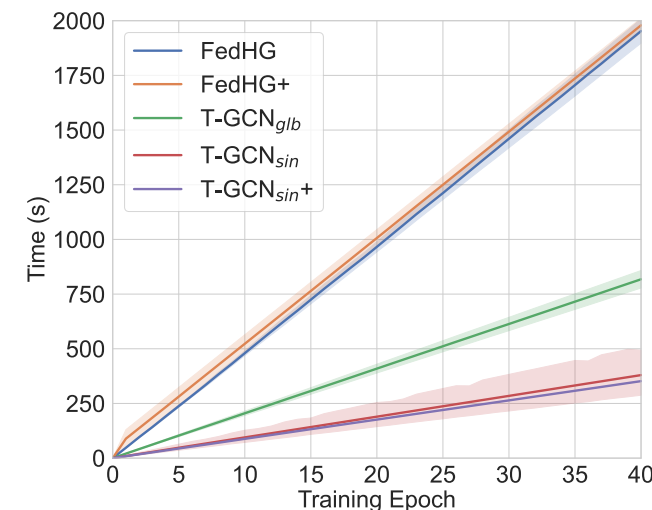
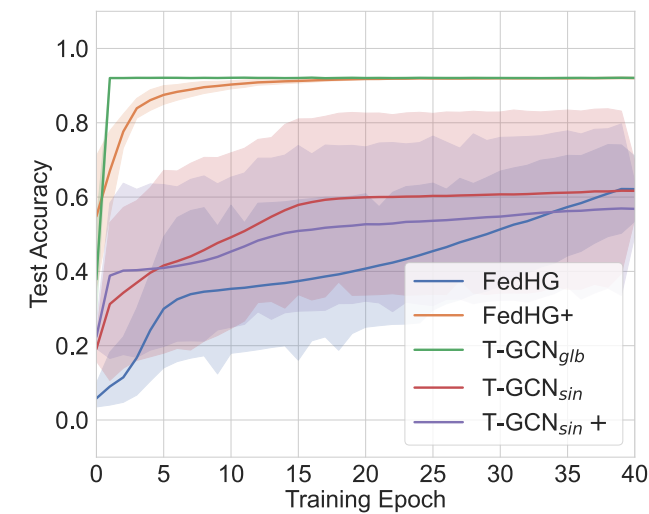
- Heterogeneous privacy requirements
- How to model heterogeneous subgraphs?
 - Expressiveness of compositional type-aware GCN
 - Generalization bound with weighted FedAvG
- How to handle incomplete neighborhoods?
 - Approximation power of type-aware NeighGen
 - Generalization bound



The diagram illustrates the framework of the proposed HNGen. It starts with an **Original graph** where node features are hidden. This leads to an **Impaired graph**. The **HNGen** module, consisting of **T-GCN** and **dHGen/fHGen** layers, processes the impaired graph to produce a **Mended graph**. The **T-GCN** layer uses a **Toy example** to show how node features are aggregated. The **dHGen/fHGen** layers use a **Toy example** to show how node features are aggregated. The final output is a **Mended graph** which is then processed by another **T-GCN** layer to produce the final output \tilde{Y} , which is used to calculate the loss \mathcal{L}^c .

Experimental Results

Model	DBLP			MIMIC-III	
	M=3	M=5	M=10	M=3	M=5
T-GCN _{sin}	0.3336 ± 0.0103	0.3360 ± 0.0306	0.3296 ± 0.0525	0.7002 ± 0.3643	0.5782 ± 0.3912
T-GCN _{sin} +	0.3325 ± 0.0024	0.3432 ± 0.0402	0.3215 ± 0.0763	0.7690 ± 0.313	0.4743 ± 0.3965
FedHG	0.3336 ± 0.0003	0.3837 ± 0.0214	0.3356 ± 0.0019	0.7401 ± 0.2165	0.6873 ± 0.2654
FedHG+	0.3343 ± 0.0006	0.4322 ± 0.0142	0.3673 ± 0.0051	0.8054 ± 0.0954	0.8492 ± 0.0565
T-GCN _{glb}	0.6419 ± 0.0010	0.7680 ± 0.0014	0.7041 ± 0.0011	0.9201 ± 0.0002	0.9206 ± 0.0004

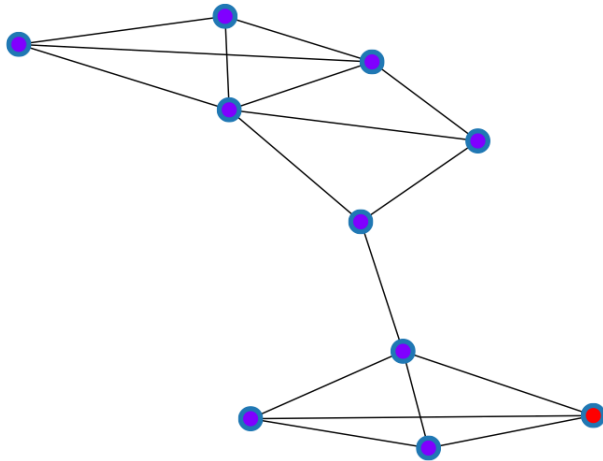


Federated Graph Classification over Non-IID Graphs [NeurIPS'21]

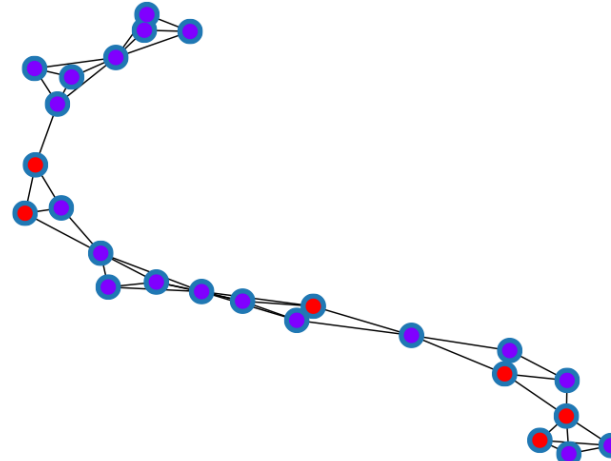
- Real scenarios of graph classification:
 - Molecules → molecular property, activity prediction, synthesis prediction, de novo drug design, etc.
 - Bioinformatics → proteins into enzymes/non-enzymes, enzymes' catalyzed levels
 - Social networks → predict the field, venue, genre, etc.
 - Geographic → predict traffic conditions, demands

Can real-world graphs from heterogeneous sources benefit the learning of each other?

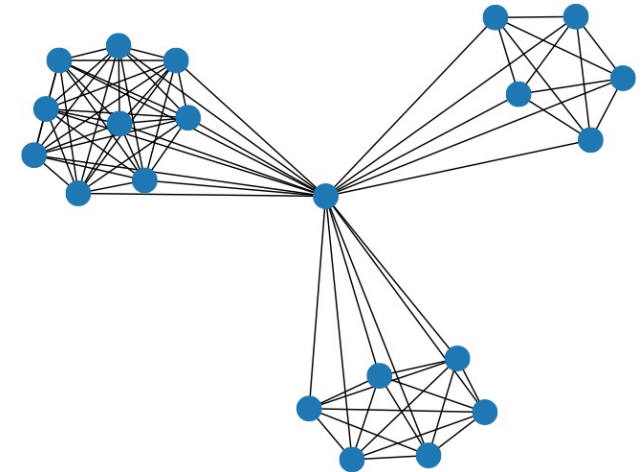
PROTEINS (bioinformatics)



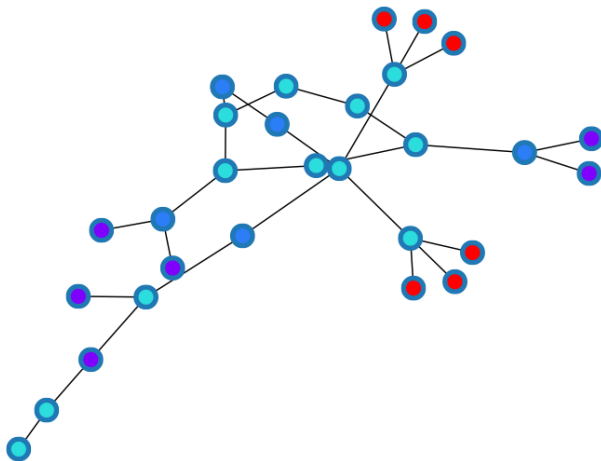
ENZYMES (bioinformatics)



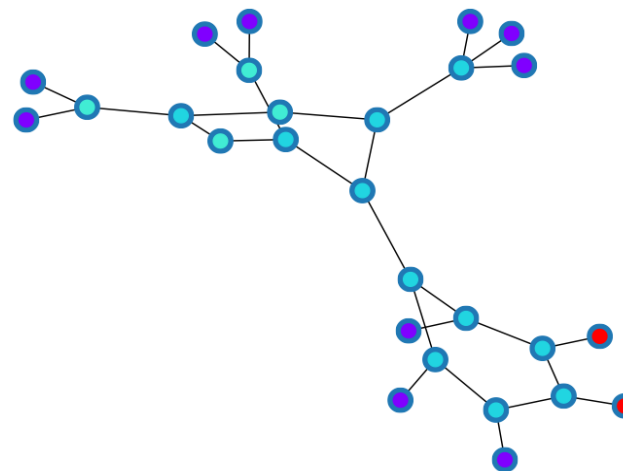
IMDB-BINARY (social networks)



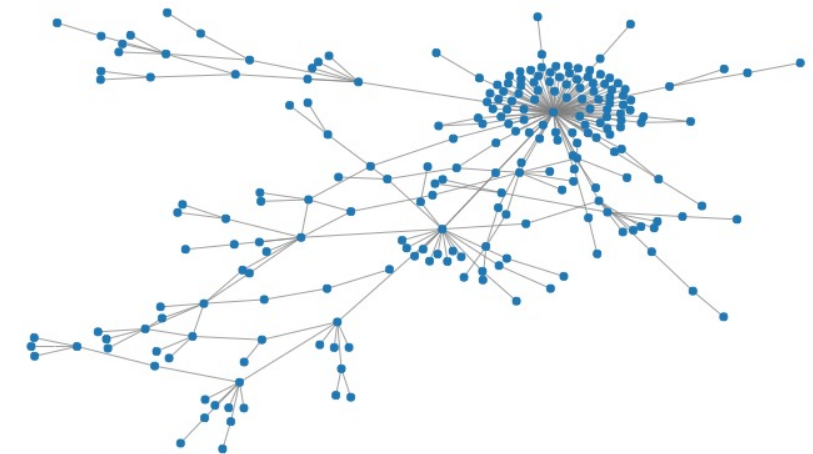
NCI1 (molecules)



DHFR (molecules)



REDDIT-BINARY (social networks)



Real-World Graphs Share General Properties

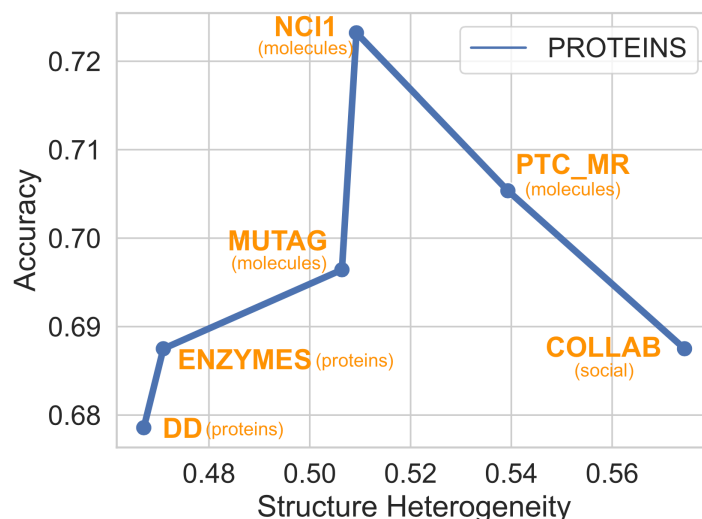
Property	kurtosis of degree distribution			avg. shortest path length			largest component size (LC, %)			clustering coefficient (CC)		
	real	random	p-value	real	random	p-value	real	random	p-value	real	random	p-value
PTC_MR (molecules)	2.1535	2.4424	0.9999	3.36	2.42	~ 0	100	82.68	~ 0	0.0095	0.1201	~ 0
ENZYMES (proteins)	3.0106	2.8243	0.0027	4.44	2.56	~ 0	98.24	97.69	0.2054	0.4516	0.1425	~ 0
IMDB-BINARY (social)	8.9262	2.2791	~ 0	1.48	1.54	~ 0	100	99.93	0.0023	0.9471	0.5187	~ 0
MSRC_21 (superpixel)	3.6959	2.9714	~ 0	4.09	2.81	~ 0	100	99.43	~ 0	0.5147	0.0655	~ 0

- Long-tail degree distribution: ENZYMES, IMDB-BINARY, MSRC_21
- Average shortest path lengths: PTC_MR, ENZYMES, MSRC_21
- Largest component size : PTC_MR, ENZYMES, IMDB-BINARY, MSRC_21
- High clustering coefficient: ENZYMES, IMDB-BINARY, MSRC_21

*Motivation for
federated graph
classification!*

Structure & Feature Heterogeneity Varies

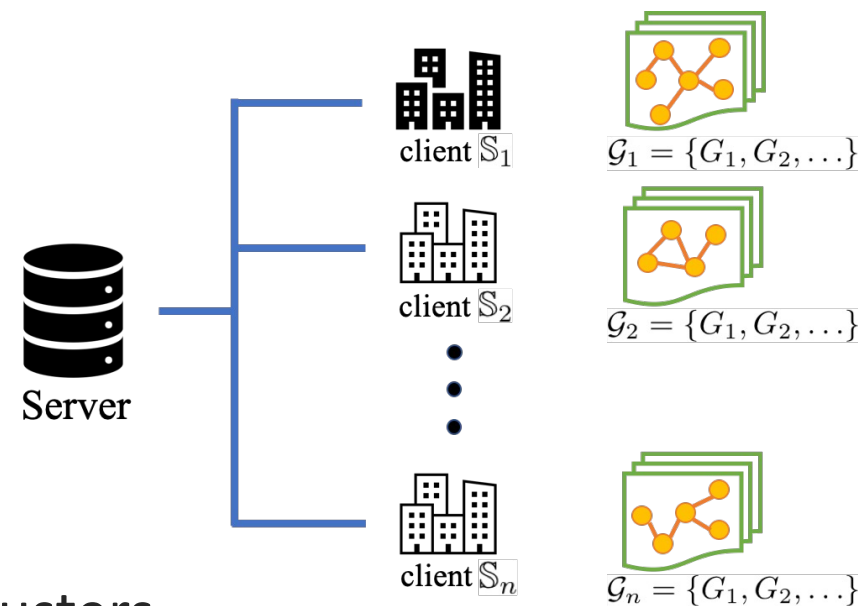
dataset	IMDB-BINARY (social)	COX2 (molecules)	COX2 (molecules) PTC_MR (molecules)	COX2 (molecules) ENZYMES (proteins)	COX2 (molecules) IMDB-BINARY (social)
avg. struc. hetero.	0.4406 (± 0.0397)	0.3246 (± 0.0145)	0.3689 (± 0.0540)	0.5082 (± 0.0399)	0.6079 (± 0.0331)
avg. feat. hetero.	0.1785 (± 0.1226)	0.0427 (± 0.0314)	0.1837 (± 0.1065)	0.1912 (± 0.1000)	0.1642 (± 0.1006)



*Careful handling
of heterogeneity
(non-IIDness) is
needed!*

GCFL -- Graph Clustered Federated Learning

- Based on CFL: a multi-task learning based framework
 - Components of GCFL:
 - One central server, a set of n local clients
 - Each client contains a set of graphs
- Dynamically cluster the clients into a set of clusters
- Perform cluster-wise FL



GCFL (GIN + CFL)

- GIN: Graph-Isomorphism Network
 - State-of-the-art simple graph-level GNN
- Sensitive to both features and structures
 - Theoretical analysis based on the Bourgain theorem
- GIN gradients can capture structure and feature heterogeneity
- The communicated gradients can capture cluster heterogeneity



GCFL (GIN + CFL)

1. General FL until the stationary point
2. At each communication round t , transmit m (*#clusters*) sets of gradients $\{\Delta\theta_i\}_{i=1}^n$; each cluster C checks:

$$\delta_{mean} = \left\| \sum_{i \in [n]} \Delta\theta_i \right\| < \epsilon_1 \quad \delta_{max} = \max(\|\Delta\theta_i\|) > \epsilon_2 > 0$$

3. If satisfied, the server calculates a cluster-wise cosine similarity matrix α_k , and perform clustering

Experiments

- **Datasets** -- 13 datasets from 3 domains, including molecules (7), proteins (3), and social networks (3)
 - *Setting 1*: single dataset, randomly distributed to a large number of clients
 - *Setting 2*: multiple datasets from single / multiple domains, one client maintains one dataset
- **3 Baselines**: self-train, FedAvg, FedProx
- **Architecture**: 3-layer GINs with pre- and post-processing MLPs

Setting 1: Single-Dataset

Table 3: Performance on the single dataset multiple client setting. We present the average accuracy and minimum gain over `self-train` on all clients, as well as the ratio of clients which get improved.

Dataset (# clients)	NCI1 (30)			PROTEINS (10)			IMDB-BINARY (10)		
Accuracy	average	min gain	ratio	average	min gain	ratio	average	min gain	ratio
<code>self-train</code>	0.6468(± 0.053)	—	—	0.7213(± 0.058)	—	—	0.7654(± 0.057)	—	—
FedAvg	0.6474(± 0.076)	-0.1333	14/30	0.7490(± 0.034)	-0.0615	6/10	0.7596(± 0.049)	-0.0800	5/10
FedProx	0.6437(± 0.072)	-0.2400	16/30	0.7556(± 0.036)	-0.0923	7/10	0.7746(± 0.048)	-0.0600	6/10
GCFL	0.7326(± 0.052)	-0.0462	26/30	0.7739(± 0.043)	-0.0545	8/10	0.8256(± 0.059)	0.0182	10/10
GCFL+	0.7422 (± 0.053)	-0.1143	28/30	0.7776 (± 0.037)	-0.0154	9/10	0.8299 (± 0.052)	0.0167	10/10

Setting 2: Multi-Dataset

Table 4: Performance on the multiple dataset multiple client setting. Metrics are the same as Table 3

Dataset (# domains)	MOLECULES (1)			BIOCHEM (2)			MIX (3)		
	average	min gain	ratio	average	min gain	ratio	average	min gain	ratio
self-train	0.7543(± 0.017)	—	—	0.7129(± 0.016)	—	—	0.7001(± 0.034)	—	—
FedAvg	0.7524(± 0.026)	-0.0132	3/7	0.6944(± 0.027)	-0.1467	4/10	0.6886(± 0.023)	-0.1233	5/13
FedProx	0.7668(± 0.032)	-0.0054	5/7	0.7053(± 0.26)	-0.1000	5/10	0.6897(± 0.026)	-0.1367	5/13
GCFL	0.7661(± 0.016)	0.0010	7/7	0.7172(± 0.019)	-0.0700	7/10	0.7056(± 0.019)	-0.1400	10/13
GCFL+	0.7745 (± 0.030)	0.0010	7/7	0.7312 (± 0.031)	-0.0300	8/10	0.7121 (± 0.021)	-0.0233	10/13

MOLECULES: 7 molecule datasets from the same domain; BIOCHEM: MOLECULES + 3 protein datasets; MIX: BIOCHEM + 3 social networks

Heterogeneity Analysis on Clusters

- Structure & feature heterogeneity within clusters

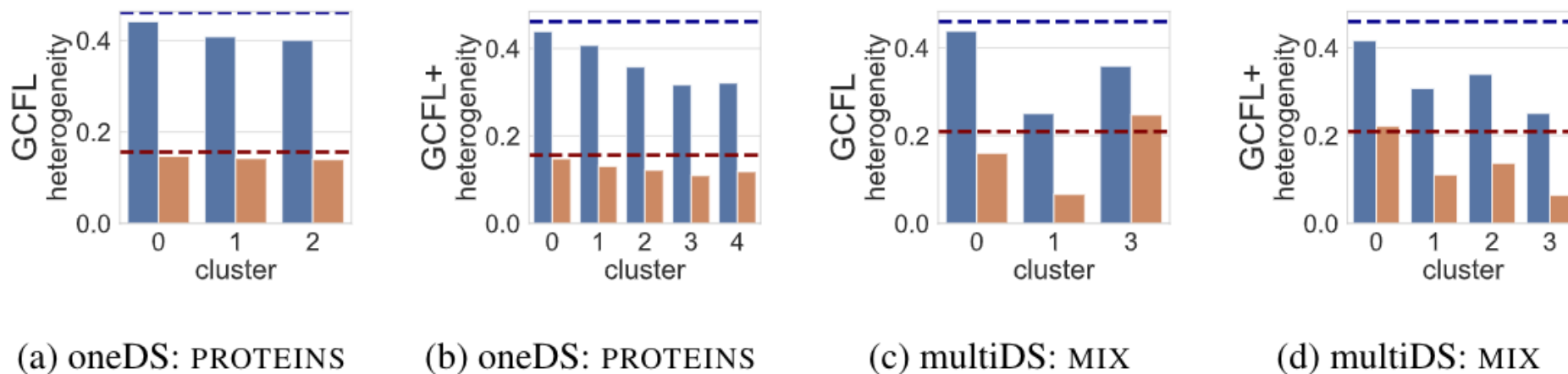
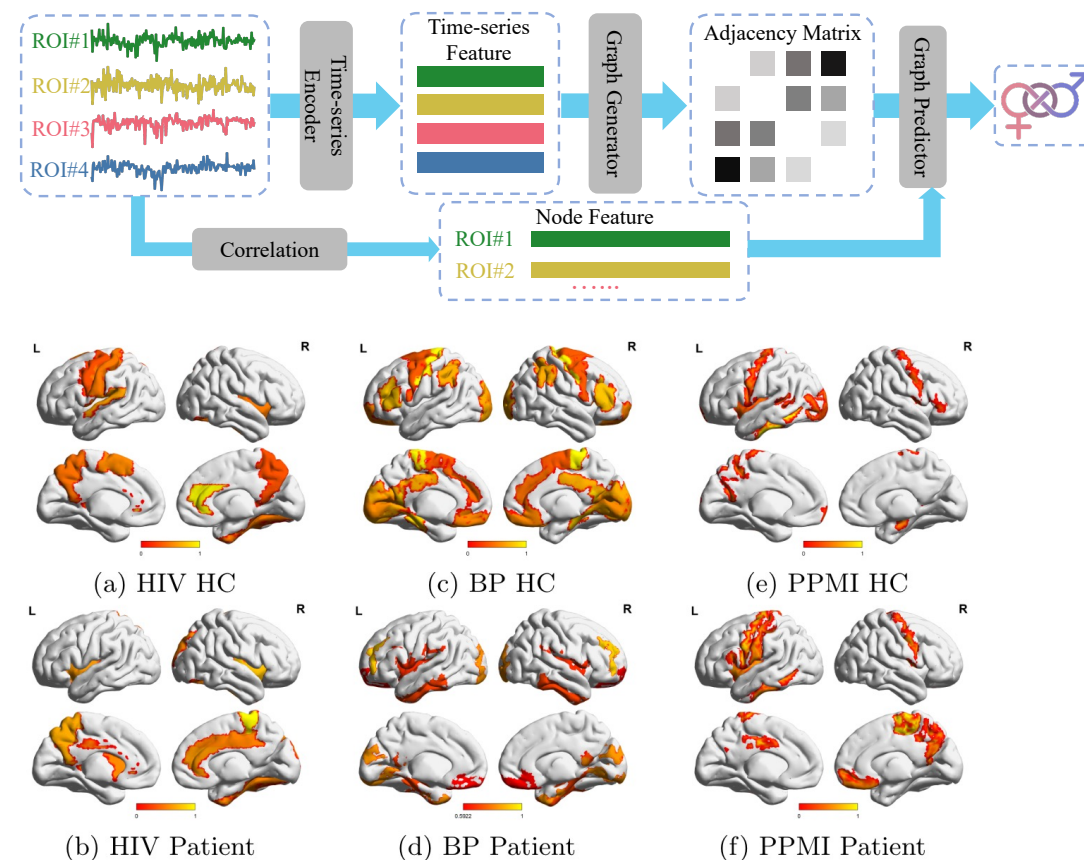


Figure 3: Structure (blue) and feature (red) heterogeneity within clusters found by GCFL and GCFL+. Dashed lines denote the heterogeneity over all clients before clustering.

Blue: structure *Yellow: feature* *Dashed line: the heterogeneity before clustering*

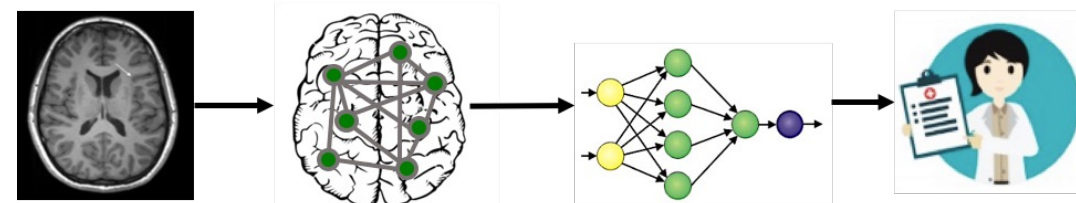
Brain Networks

- FBNetGen (MIDL'22)
 - End-to-end brain network generation + gender prediction + interpretation
- BrainNNExplainer (MICCAI'22)
 - GNNExplainer applied on brain networks for HIV, BP and PPMI
- MBNN (EMBC'22)
 - Multiview-GNN applied on brain networks for HIV, BP and PPMI
- PGB (KDD'22, CHIL'23)
 - Pre-training GNNs for brain network analysis



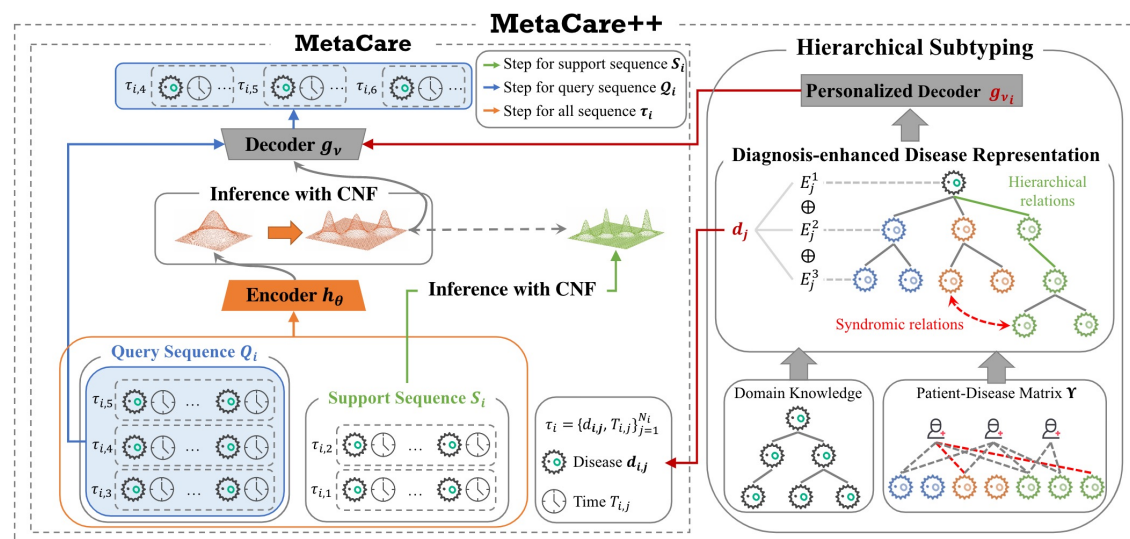
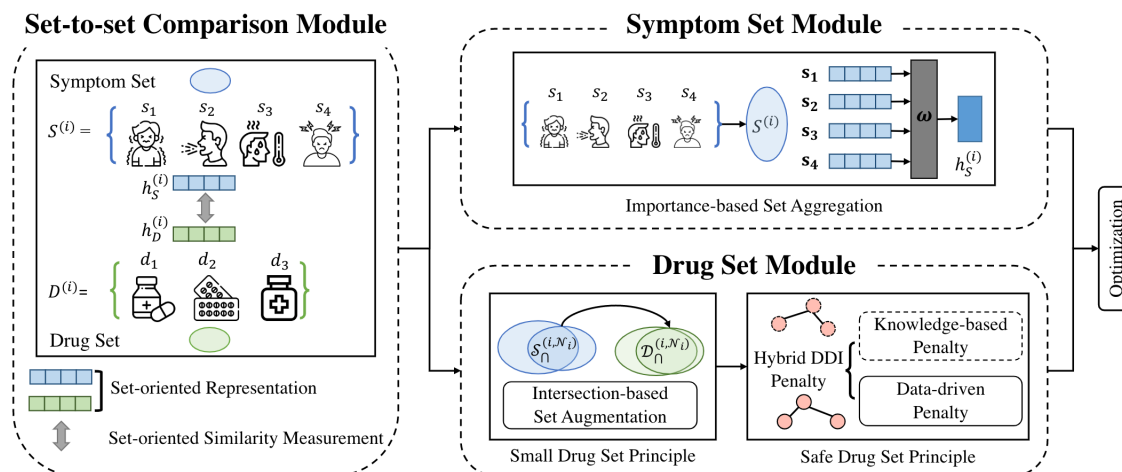
Brain Networks + FedGraph

- Motivation: limited data in local studies (amount and access)
- Scenarios
 - Single disease multiple cohorts:
e.g., Parkinson's, depression
 - Single cohort multiple diseases:
e.g., Parkinson's, Alzheimer's; PTSD, depression
- Unique challenges (preliminaries)
 - Node system alignment: e.g., PGB (CHIL'23)
 - Cross task/dataset transferability: e.g., EGI (NeurIPS'21)
 - Self-supervised GNN training: e.g., NWR-GAE (ICLR'22)



EHR Networks

- 4SDrug (KDD'22)
 - Set-to-set (symptom-to-drug) recommendation
 - Important set of symptoms
 - Small and safe set of drugs
- MetaCare (SIGIR'22)
 - Meta-learning for rare diseases and infrequent patients
 - External disease relationships + data-driven disease complications
- CACHE (ML4H'22 Best Paper)
 - Hypergraph modeling of medical codes
 - Counterfactual and factual reasoning for interpretability



EHR Networks + FedGraph

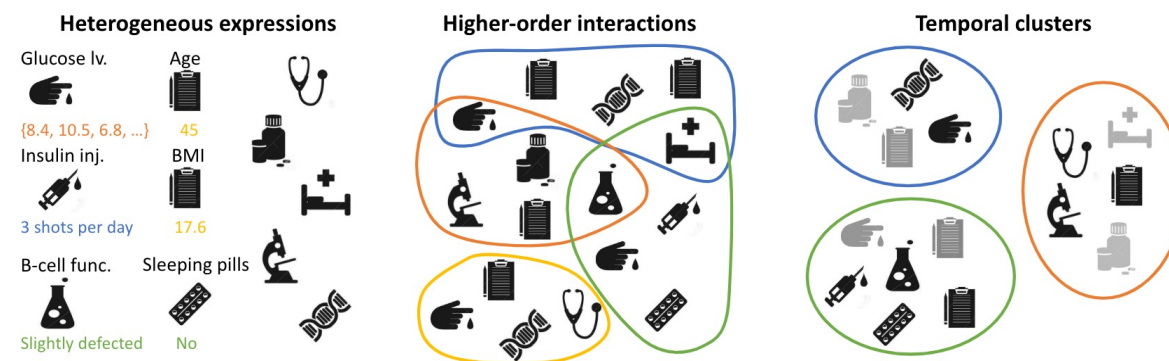
- Motivation: limited diversity in single studies (patient and disease)

- Scenarios:

- Multi-modality data beyond EHR?
- Disease phenotyping/subtyping?
- Multi-task learning?

- Unique challenges:

- Heterogeneous expressions: e.g., HGB (TKDE'20)
- Higher-order interactions: e.g., HypEHR (AMIA'22)
- Temporal clusters: e.g., DyHINE (WSDM'21), EvoNet (WSDM'21)



Suggestions/Questions

