Inductive Attributed Community Search: to Learn Communities across Graphs

Shuheng Fang[†], Kangfei Zhao[∮], Yu Rong[*], Zhixun Li[†], Jeffrey Xu Yu[†]

The Chinese University of Hong Kong[†]

Beijing Institute of Technology[∮]

Alibaba DAMO Academy[*]

VLDB 2024

Background: Attributed Community Search

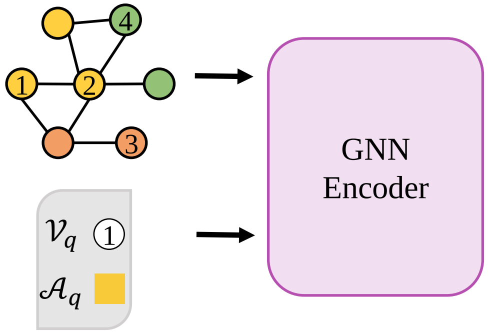
- Community Search (CS)
  - For a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, given a node set $\mathcal{V}_q \subseteq \mathcal{V}$as a query q
  - Find the query-dependent community $C_q \subseteq \mathcal{V}$, where the nodes in C_q are intensively intra-connected

- Attributed Community Search (ACS)
  - Satisfy both structure cohesiveness and attributes homogeneity for a given query that consists of query nodes and query attributes.

- Real applications
  - Social network analysis
  - Recommendation systems
  - Bioinformatics and fraud detection

Background: Learning-based Community Search Approaches

- GNN-based method: recasting the community membership determination to a classification task
- AQD-GNN/ICS-GNN
  - Their trained models are tailored for specific graph/community
- ICS-GNN/COCLEP/CommunityAF/CGNP
  - Only support single-node query
  - COCLEP & CommunityAF have a limited inductive ability as they rely on the natural generalization of GNN
  - CGNP utilizes meta-learning approach and has inductive ability



Background: Learning-based Community Search Approaches

Table 1: Learning-based Community Search Approaches

Approaches	Single-node Query	Multi-node Query	Attributed Query	Induction
AQD-GNN [29]	✔	✔	✔	✘
ICS-GNN [21]	✔	✘	✘	✘
CommunityAF [11]	✔	✘	✘	✔✘
COCLEP [35]	✔	✘	✘	✔✘
CGNP [17]	✔	✘	✘	✔
IACS (Ours)	✔	✔	✔	✔

Problem Statement

- Attributed Community Search (ACS)
  - For an attributed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{A})$, given a query $q = (\mathcal{V}_q, \mathcal{A}_q)$, where $\mathcal{V}_q \subseteq \mathcal{V}$ and $\mathcal{A}_q \subseteq \mathcal{A}$
  - Find the query-dependent community $C_q \subseteq \mathcal{V}$, where the nodes in C_q are intensively intra-connected and the attributes are similar

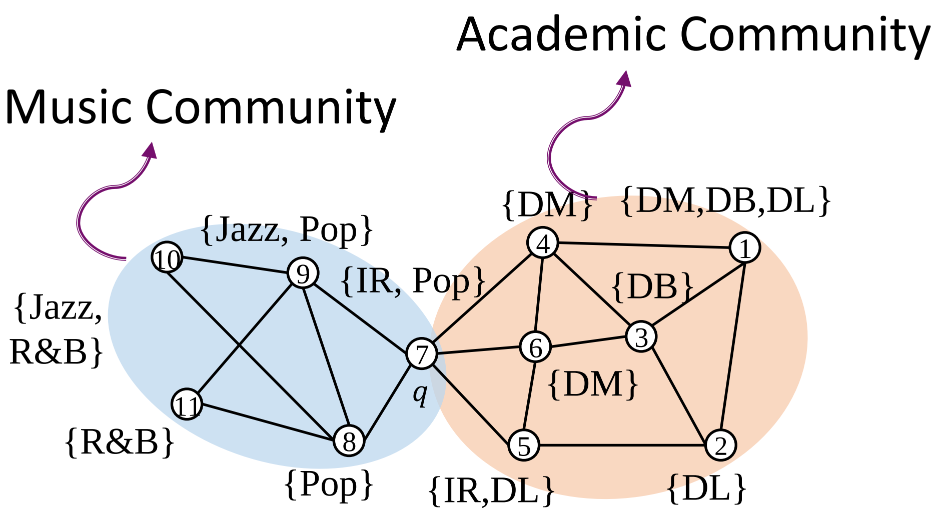
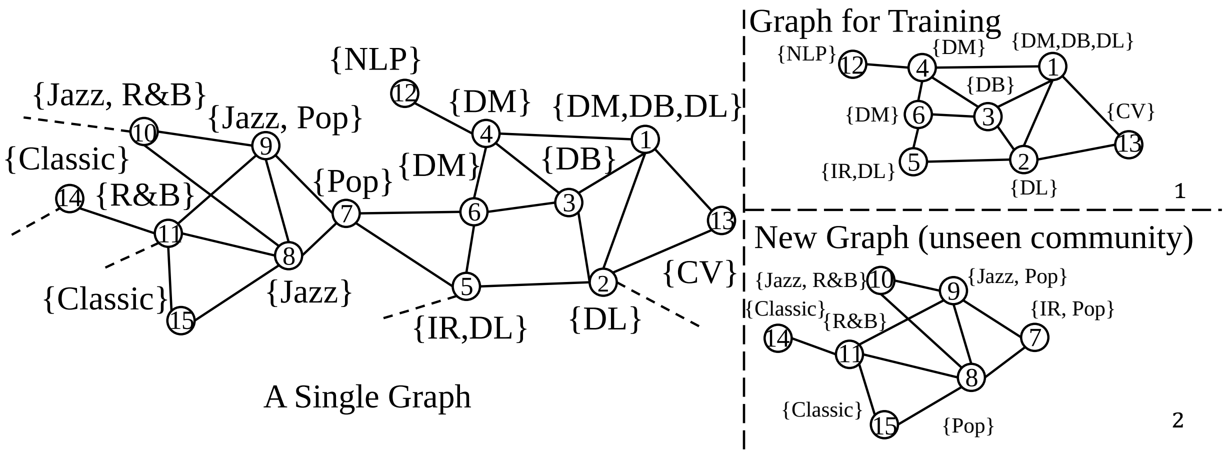


Fig 1. query-dependent community with different attributes.

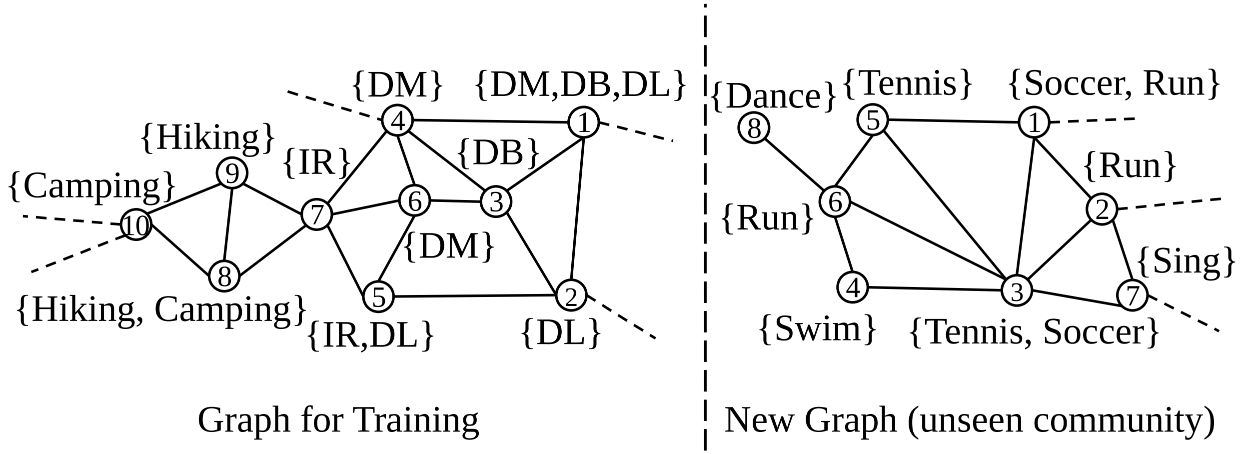
Problem Statement

Empower the model to generalize and adapt to new communities and graphs by inductive learning

- For new communities
  - Queries from $\{C_{q_1}, \ldots, C_{q_i}\}$ in graph $\mathcal{G}$ for training
  - Queries q^* from a new community C_{q^*} for test
  - i.e., $C_{q_1} \cap C_{q^*} = \emptyset, \ldots, C_{q_i} \cap C_{q^*}$
- For new graphs
  - Queries from graph $\mathcal{G}$ for training
  - Queries from new graph $\mathcal{G}^*$ for test

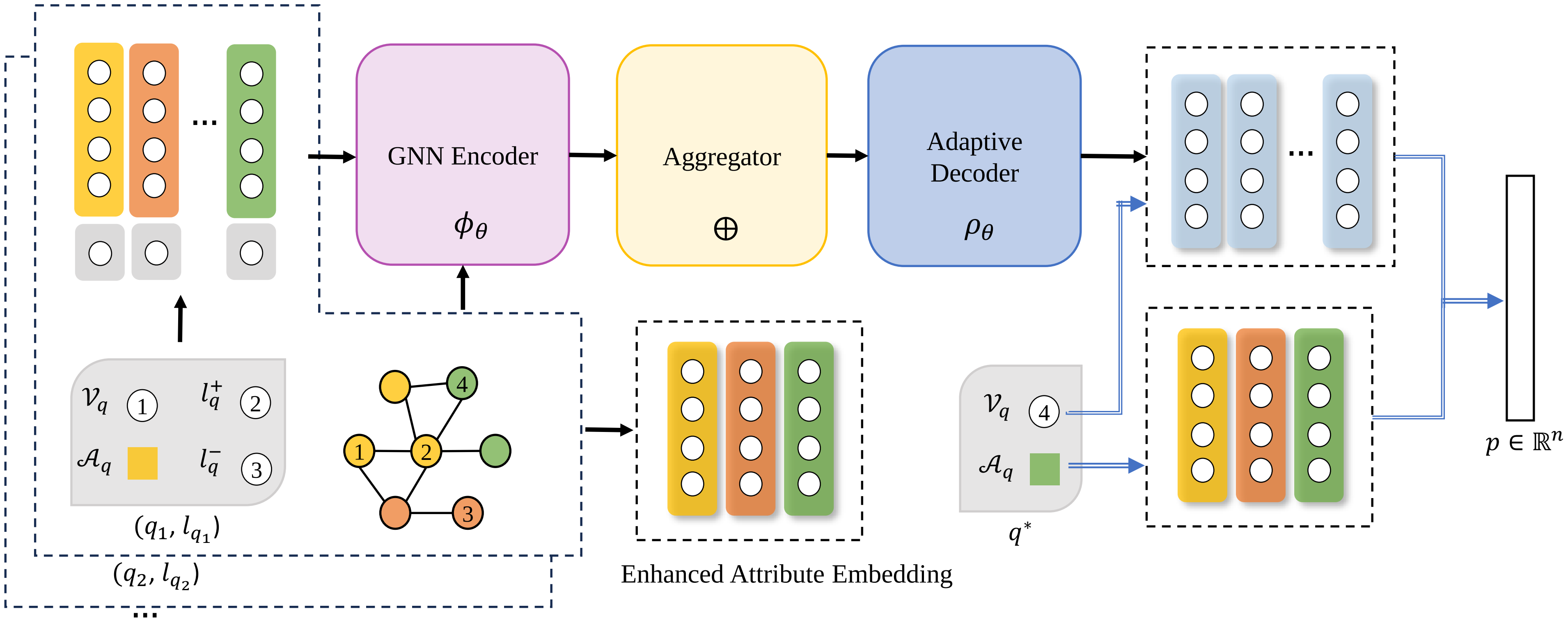


(a) Inductive Setting for Communities



(b) Inductive Setting for Graphs

IACS Architecture



IACS

$$p(y_{q^*}|q^*,\mathcal{T}) = \rho_\theta\left(q^*, \bigoplus_{(q,l_q)\in(Q,L)} \phi_\theta(q,l_q)\right)$$

GNN Encoder $\phi_\theta\ (q,l_q)$

- k-th layer aggregate: $a^{(k)}(v) \leftarrow f_{\mathcal{A}}^{(k)}(\{h^{(k-1)}(u)\ |u \in N(v)\})$
- k-th layer combine: $h^{(k)}(v) \leftarrow f_{\mathcal{C}}^{(k)}(h^{(k-1)}(v), a^{(k)}(v))$

Aggregator $\oplus$

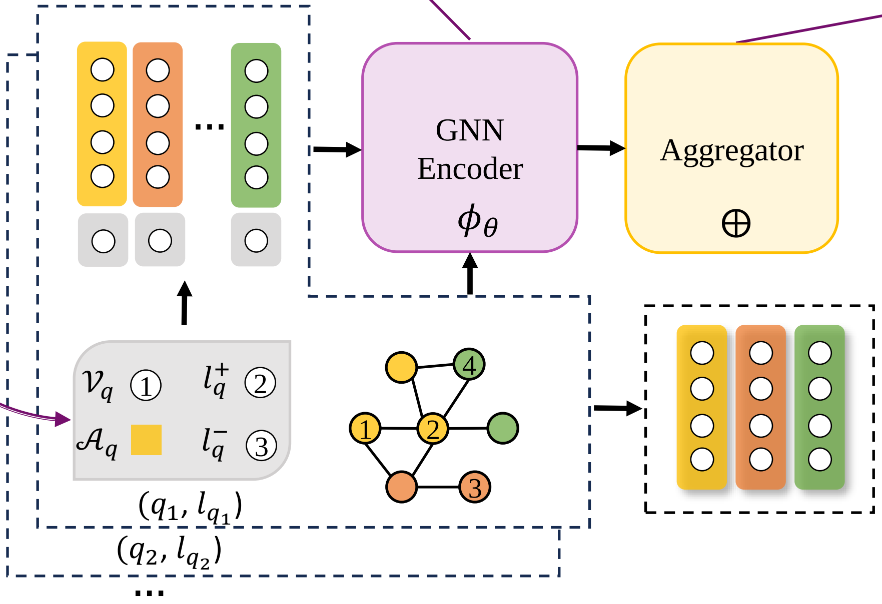
Permutation invariant operator,

average: $H = \frac{1}{|Q|}\sum_{q\in Q} H_q$

Inputs

$h^{(0)}(v) = [I_l(v)||e(v)]$,

$$I_l(v) = \begin{cases} 1 & v \in l_q^+ \cup \{\mathcal{V}_q\} \\ 0 & \text{otherwise} \end{cases}$$

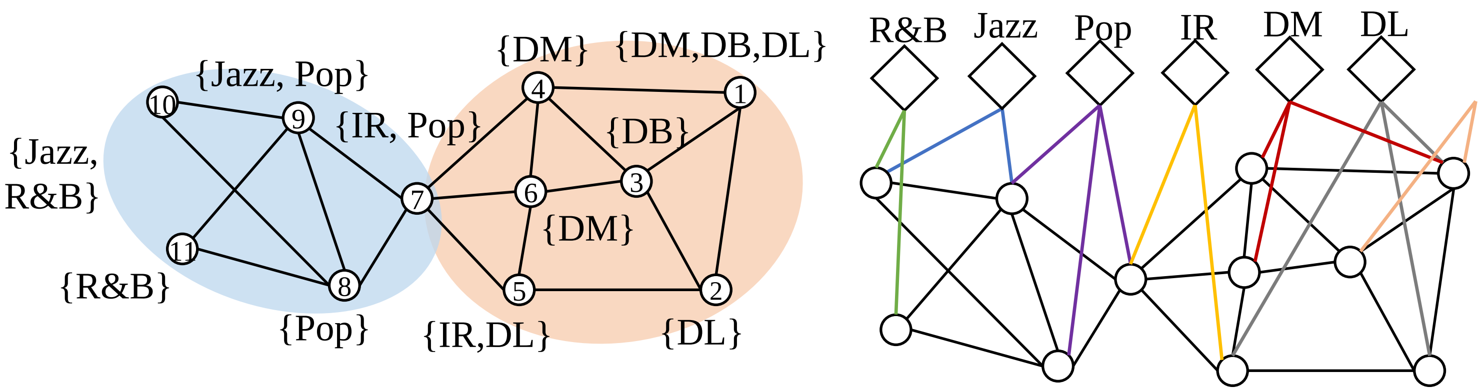


Enhanced Attribute Encoding

Construct attribute-augmented graph

- $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{A}) \rightarrow \mathcal{G}_{\mathcal{A}} = (\mathcal{V} \cup \mathcal{V}_{\mathcal{A}}, \mathcal{E} \cup \mathcal{E}_{\mathcal{A}})$
- Use a scalable, task-independent graph embedding algorithm, ProNE
- Pretrain a node embedding for the attributed-augmented graph $\mathcal{G}_{\mathcal{A}}$

$$e(v) = \sum_{a \in \mathcal{A}(v)} e_a$$



IACS

Gating mechanism

$\delta = \text{sigmod}(W_\delta H), \epsilon = W_\epsilon H,$

$\hat{\gamma} = \gamma \odot \delta + \epsilon \odot (1 - \delta), \hat{\beta} = \beta \odot \delta + \epsilon \odot (1 - \delta),$

$\widehat{H} = \hat{\gamma} \odot H + \hat{\beta}$

Adaptive Decoder $\rho_\theta(q^*, H)$

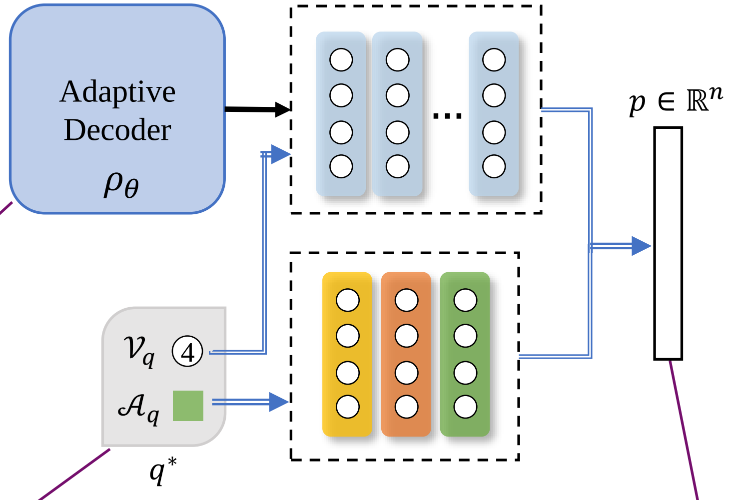
- Feature-wise Linear Modulation (FiLM)

$$\gamma = W_\gamma H,\ \beta = W_\beta H,$$
$$\widehat{H} = \gamma \odot H + \beta$$

- Concatenate query node embedding and query attribute embedding

$$e_{\mathcal{V}_{q^*}} = \frac{1}{|\mathcal{V}_{q^*}|} \sum_{v \in \mathcal{V}_{q^*}} \widehat{H}(v), e_{\mathcal{A}_{q^*}} = \frac{1}{|\mathcal{A}_{q^*}|} \sum_{a \in \mathcal{A}_{q^*}} e_a$$
$$e_{q^*} = \text{MLP}(e_{\mathcal{V}_{q^*}} || e_{\mathcal{A}_{q^*}})$$

- Inner Product Decoder: $p\left(\hat{l}_{q^*} \middle| q^*, \mathcal{T}\right) = \text{sigmoid}(< e_{q^*}, \widehat{H} >)$



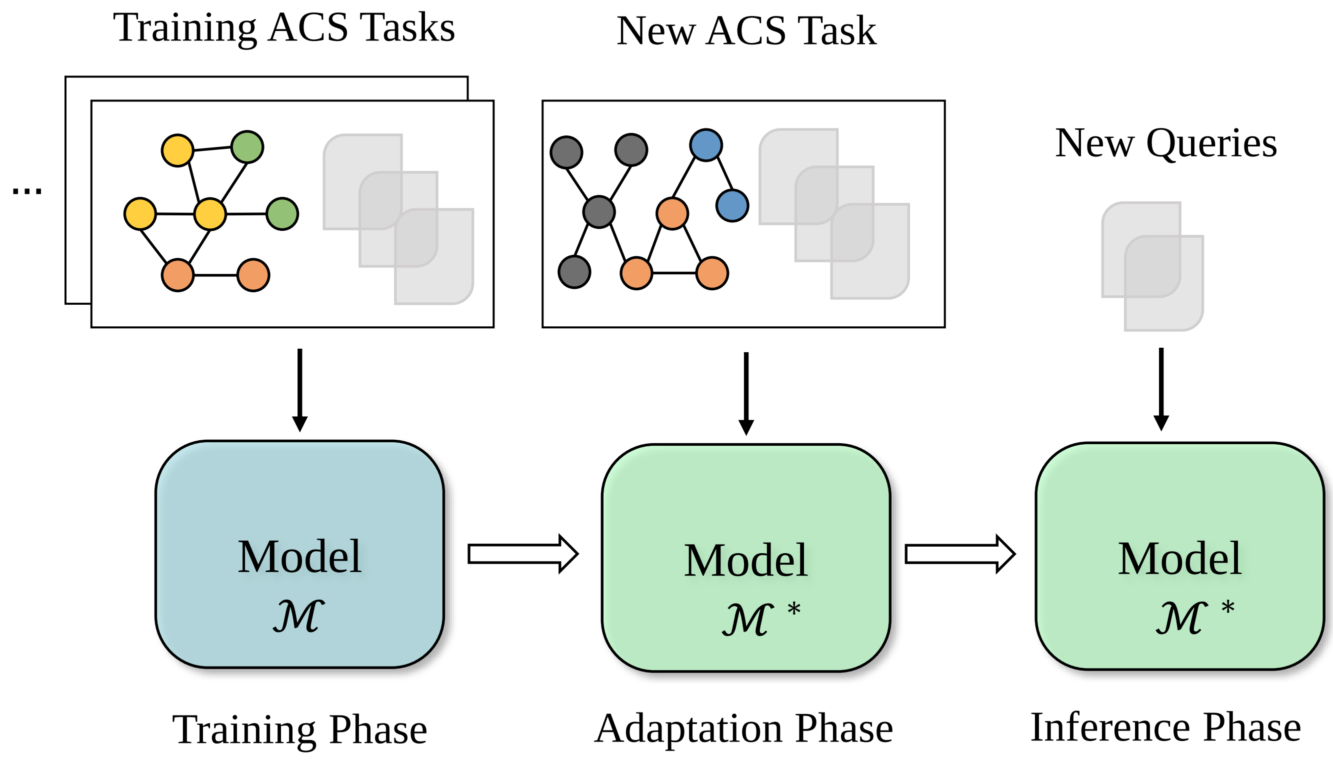
A new ACS query

$$q^* = (\mathcal{V}_{q^*}, \mathcal{A}_{q^*})$$

Binary Cross Entropy loss

$$\mathcal{L} = \sum_{\mathcal{T}_i \in \mathcal{D}} \sum_{(q, l_q) \in (\mathcal{Q}_i, L_i)} -\log p\left(\widehat{l_q} \middle| q, \mathcal{T}_i\right)$$
$$= \sum_{\mathcal{T}_i \in \mathcal{D}} \sum_{(q, l_q) \in (\mathcal{Q}_i, L_i)} \left(-\sum_{v^+ \in l_q^+} \log(\hat{y}(v^+)) - \sum_{v^- \in l_q^-} \log(1 - \hat{y}(v^-))\right)$$

IACS Workflow



Experimental Studies: Setup

- Model
  - encoder: 3-layer GCN, GraphSAGE, GIN, GAT.
  - decoder: FiLM + Inner product, FiLM with gating mechanism + Inner product, inner product
- Baselines
  - 3 algorithmic approaches
  - 3 supervised-learning based approaches
  - 2 meta-learning approaches

Table 3: The Profiles of Dataset

Dataset	$\|\mathcal{G}\|$	$\|\mathcal{V}\|$	$\|\mathcal{E}\|$	$\|\mathcal{A}\|$	$\|C\|$	graph des.	attribute des.	community des.	# tasks
Arxiv [54]	1	169,343	1,166,243	N/A	40	paper citation	NA	research topics	1,000
Amazon2M [13]	1	2,449,029	61,859,140	N/A	47	product co-purchasing	NA	product categories	5,000
Cora [57]	1	2,708	5,429	1,433	7	paper citation	paper keywords	research topics	192
Citeseer [57]	1	3,327	4,732	3,703	6	paper citation	paper keywords	research topics	192
Reddit [24]	1	232,965	114,615,892	1,164	50	post co-comment	synthetic	post categories	1,000
Facebook [33]	10	4,039	88,234	2,281	193	social friendship	user profiles	friend circles	10
Twitter [33]	973	81,306	1,768,149	512,985	4,065	social friendship	user profiles	friend circles	973

Experiential Studies: Effectiveness

Table 4: Overall Performance on Non-Attributed CS (%)

Dataset	Approach	4-shot Pre	4-shot Rec	4-shot F1	8-shot Pre	8-shot Rec	8-shot F1
Arxiv	*CTC*	$54.23_{\pm0.53}$	$2.16_{\pm0.04}$	$4.15_{\pm0.09}$	$54.04_{\pm0.72}$	$2.16_{\pm0.05}$	$4.15_{\pm0.09}$
	ICS-GNN	$62.72_{\pm0.26}$	$21.09_{\pm0.07}$	$31.57_{\pm0.10}$	$62.53_{\pm0.36}$	$21.12_{\pm0.05}$	$31.57_{\pm0.08}$
	QD-GNN	$59.97_{\pm0.41}$	$83.60_{\pm1.18}$	$69.84_{\pm0.47}$	$58.91_{\pm0.29}$	$89.62_{\pm1.14}$	$71.09_{\pm0.29}$
	Supervise	$67.99_{\pm0.33}$	$69.78_{\pm1.49}$	$68.87_{\pm0.86}$	$69.09_{\pm0.39}$	$74.29_{\pm0.98}$	$71.60_{\pm0.38}$
	MAML	$63.51_{\pm1.07}$	$60.25_{\pm2.50}$	$61.81_{\pm1.42}$	$62.77_{\pm0.72}$	$60.34_{\pm3.64}$	$61.48_{\pm1.82}$
	FeatTrans	$65.35_{\pm0.64}$	$55.18_{\pm1.81}$	$59.81_{\pm0.88}$	$64.18_{\pm0.69}$	$55.42_{\pm1.09}$	$59.47_{\pm0.74}$
	IACS	$63.65_{\pm0.62}$	$89.26_{\pm1.05}$	**$74.31_{\pm0.37}$**	$64.14_{\pm0.49}$	$90.21_{\pm1.31}$	**$74.97_{\pm0.31}$**
	IACS-G	$59.75_{\pm0.42}$	$97.99_{\pm0.76}$	$74.23_{\pm0.13}$	$65.06_{\pm0.81}$	$88.12_{\pm1.65}$	$74.84_{\pm0.36}$
	IACS-P	$61.99_{\pm2.56}$	$92.63_{\pm5.77}$	$74.12_{\pm0.21}$	$65.45_{\pm0.42}$	$87.07_{\pm0.53}$	$74.72_{\pm0.24}$
Amazon2M	*CTC*	$80.30_{\pm0.35}$	$4.06_{\pm0.02}$	$7.73_{\pm0.04}$	$80.27_{\pm0.27}$	$4.06_{\pm0.01}$	$7.73_{\pm0.02}$
	ICS-GNN	$79.50_{\pm0.27}$	$6.55_{\pm0.01}$	$12.11_{\pm0.02}$	$79.63_{\pm0.29}$	$6.55_{\pm0.02}$	$12.11_{\pm0.03}$
	QD-GNN	$75.46_{\pm0.33}$	$95.15_{\pm0.53}$	$84.17_{\pm0.04}$	$75.33_{\pm0.26}$	$96.68_{\pm0.13}$	$84.67_{\pm0.21}$
	Supervise	$83.86_{\pm0.09}$	$77.07_{\pm0.44}$	$80.32_{\pm0.25}$	$84.46_{\pm0.35}$	$80.18_{\pm0.52}$	$82.27_{\pm0.29}$
	MAML	$78.48_{\pm1.62}$	$65.83_{\pm8.70}$	$71.38_{\pm5.59}$	$79.13_{\pm0.88}$	$62.38_{\pm4.76}$	$69.66_{\pm2.83}$
	FeatTrans	$78.41_{\pm0.92}$	$57.89_{\pm1.39}$	$66.60_{\pm1.14}$	$78.69_{\pm0.34}$	$57.18_{\pm1.22}$	$66.22_{\pm0.72}$
	IACS	$80.52_{\pm0.34}$	$93.42_{\pm0.83}$	$86.48_{\pm0.22}$	$81.44_{\pm0.75}$	$93.34_{\pm1.07}$	$86.97_{\pm0.21}$
	IACS-G	$79.92_{\pm0.31}$	$94.25_{\pm0.93}$	$86.49_{\pm0.21}$	$80.49_{\pm0.16}$	$94.60_{\pm0.52}$	$86.98_{\pm0.24}$
	IACS-P	$79.63_{\pm0.88}$	$94.77_{\pm1.09}$	**$86.54_{\pm0.29}$**	$80.62_{\pm0.81}$	$94.86_{\pm0.74}$	**$87.16_{\pm0.26}$**

Non-attributed CS

a) IACS models consistently outperform all the baselines.
b) The superiority of IACS is primarily evident in its significant improvement in recall (+1.28% compared to the best baseline) while maintaining a relatively high precision (59.75% ~ 81.44%).

Experiential Studies: Effectiveness

Table 5: Overall Performance on ACS in Single Graph (%)

Dataset	Approach	4-shot Pre	4-shot Rec	4-shot F1	8-shot Pre	8-shot Rec	8-shot F1
Citeseer	*ATC*	$58.99_{\pm 0.87}$	$5.01_{\pm 0.17}$	$9.24_{\pm 0.29}$	$57.83_{\pm 0.36}$	$4.97_{\pm 0.25}$	$9.16_{\pm 0.42}$
	ACQ	$70.59_{\pm 2.14}$	$6.97_{\pm 1.32}$	$12.66_{\pm 2.19}$	$69.15_{\pm 1.43}$	$6.79_{\pm 0.99}$	$12.36_{\pm 1.64}$
	AQD-GNN	$52.70_{\pm 1.04}$	$84.29_{\pm 7.59}$	$64.77_{\pm 2.78}$	$52.26_{\pm 2.06}$	$85.15_{\pm 5.57}$	$64.74_{\pm 3.01}$
	Supervise	$60.45_{\pm 2.00}$	$63.20_{\pm 1.53}$	$61.79_{\pm 1.67}$	$62.30_{\pm 1.88}$	$66.74_{\pm 0.84}$	$64.43_{\pm 1.16}$
	MAML	$55.53_{\pm 1.61}$	$43.18_{\pm 4.27}$	$48.46_{\pm 2.45}$	$56.15_{\pm 0.94}$	$45.02_{\pm 3.05}$	$49.93_{\pm 1.94}$
	FeatTrans	$58.67_{\pm 2.54}$	$37.97_{\pm 1.38}$	$46.08_{\pm 1.51}$	$58.44_{\pm 1.74}$	$39.86_{\pm 2.17}$	$47.38_{\pm 1.91}$
	IACS	$64.74_{\pm 1.55}$	$71.15_{\pm 1.45}$	$67.78_{\pm 0.94}$	$67.06_{\pm 1.45}$	$71.48_{\pm 1.84}$	$69.19_{\pm 1.41}$
	IACS-G	$65.75_{\pm 0.54}$	$70.12_{\pm 1.33}$	$\mathbf{67.86}_{\pm 0.56}$	$67.48_{\pm 1.62}$	$71.90_{\pm 1.98}$	$\mathbf{69.59}_{\pm 0.87}$
	IACS-P	$65.52_{\pm 1.15}$	$70.37_{\pm 1.43}$	$67.84_{\pm 0.66}$	$67.25_{\pm 1.86}$	$72.08_{\pm 0.74}$	$69.57_{\pm 0.98}$
Reddit	*ATC*	$82.84_{\pm 0.77}$	$39.15_{\pm 1.68}$	$53.16_{\pm 1.63}$	$83.73_{\pm 0.87}$	$39.00_{\pm 0.91}$	$53.21_{\pm 0.79}$
	ACQ	$97.74_{\pm 0.27}$	$22.82_{\pm 1.02}$	$36.99_{\pm 1.35}$	$98.16_{\pm 0.03}$	$22.49_{\pm 1.20}$	$36.58_{\pm 1.59}$
	AQD-GNN	$85.88_{\pm 1.63}$	$89.54_{\pm 1.89}$	$\mathbf{87.67}_{\pm 1.38}$	$85.51_{\pm 1.73}$	$92.20_{\pm 1.39}$	$\mathbf{88.73}_{\pm 1.42}$
	Supervise	$86.16_{\pm 1.38}$	$78.14_{\pm 1.93}$	$81.95_{\pm 1.51}$	$87.14_{\pm 0.95}$	$80.15_{\pm 0.49}$	$83.50_{\pm 0.49}$
	MAML	$88.30_{\pm 0.78}$	$64.46_{\pm 3.15}$	$74.66_{\pm 2.27}$	OOM	OOM	OOM
	FeatTrans	$87.76_{\pm 2.41}$	$34.78_{\pm 3.22}$	$49.72_{\pm 3.34}$	$88.07_{\pm 0.59}$	$36.46_{\pm 2.43}$	$51.53_{\pm 2.40}$
	IACS	$84.03_{\pm 1.20}$	$84.11_{\pm 3.84}$	$84.01_{\pm 1.26}$	$83.50_{\pm 1.69}$	$85.07_{\pm 4.36}$	$84.20_{\pm 1.49}$
	IACS-G	$83.81_{\pm 1.73}$	$85.33_{\pm 2.23}$	$84.54_{\pm 1.14}$	$86.07_{\pm 0.86}$	$84.01_{\pm 2.63}$	$85.00_{\pm 1.11}$
	IACS-P	$84.85_{\pm 1.28}$	$83.90_{\pm 2.53}$	$84.35_{\pm 1.31}$	$85.89_{\pm 1.84}$	$83.21_{\pm 1.72}$	$84.51_{\pm 1.33}$

Table 6: Overall Performance on ACS in Multiple Graphs (%)

Dataset	Approach	4-shot Pre	4-shot Rec	4-shot F1	8-shot Pre	8-shot Rec	8-shot F1
Facebook	*ATC*	$60.23_{\pm 5.10}$	$11.99_{\pm 0.88}$	$19.97_{\pm 1.26}$	$41.14_{\pm 4.18}$	$11.11_{\pm 3.27}$	$17.22_{\pm 3.71}$
	ACQ	$38.86_{\pm 3.52}$	$66.92_{\pm 5.79}$	$48.90_{\pm 1.92}$	$40.60_{\pm 3.06}$	$64.00_{\pm 3.33}$	$49.65_{\pm 3.07}$
	AQD-GNN	$37.71_{\pm 1.65}$	$96.70_{\pm 5.93}$	$54.26_{\pm 2.59}$	$36.71_{\pm 1.03}$	$96.29_{\pm 4.68}$	$53.14_{\pm 1.75}$
	Supervise	$59.32_{\pm 1.22}$	$79.61_{\pm 5.37}$	$67.95_{\pm 2.73}$	$64.34_{\pm 1.63}$	$83.38_{\pm 3.20}$	$72.59_{\pm 1.08}$
	MAML	$47.06_{\pm 6.12}$	$89.04_{\pm 3.86}$	$59.85_{\pm 8.12}$	$46.12_{\pm 3.30}$	$73.30_{\pm 5.89}$	$56.44_{\pm 2.45}$
	FeatTrans	$50.11_{\pm 6.62}$	$68.34_{\pm 8.66}$	$56.84_{\pm 4.28}$	$50.82_{\pm 1.16}$	$59.77_{\pm 6.87}$	$72.16_{\pm 3.40}$
	IACS	$85.61_{\pm 2.16}$	$79.55_{\pm 4.13}$	$\mathbf{78.09}_{\pm 4.09}$	$66.13_{\pm 1.62}$	$84.33_{\pm 3.80}$	$74.08_{\pm 1.34}$
	IACS-G	$85.75_{\pm 2.27}$	$81.31_{\pm 4.51}$	$77.42_{\pm 3.82}$	$64.87_{\pm 1.67}$	$88.17_{\pm 3.04}$	$74.72_{\pm 1.64}$
	IACS-P	$81.82_{\pm 4.42}$	$80.79_{\pm 2.69}$	$77.37_{\pm 4.62}$	$65.92_{\pm 4.20}$	$87.36_{\pm 1.66}$	$\mathbf{75.05}_{\pm 2.31}$
Twitter2Facebook	*ATC*	$77.92_{\pm 6.75}$	$12.88_{\pm 0.74}$	$22.08_{\pm 1.04}$	$70.30_{\pm 9.71}$	$11.25_{\pm 1.91}$	$19.35_{\pm 3.05}$
	ACQ	$68.89_{\pm 0.81}$	$37.21_{\pm 8.31}$	$49.51_{\pm 4.76}$	$20.38_{\pm 0.58}$	$44.58_{\pm 3.51}$	$28.22_{\pm 0.47}$
	AQD-GNN	$37.49_{\pm 0.49}$	$96.81_{\pm 3.49}$	$37.49_{\pm 1.03}$	$37.61_{\pm 3.45}$	$95.10_{\pm 8.97}$	$53.84_{\pm 4.54}$
	Supervise	$58.12_{\pm 4.20}$	$80.28_{\pm 8.12}$	$58.12_{\pm 5.27}$	$62.32_{\pm 4.41}$	$81.44_{\pm 4.99}$	$70.55_{\pm 4.15}$
	MAML	$38.12_{\pm 0.42}$	$97.57_{\pm 1.67}$	$38.12_{\pm 0.37}$	$37.97_{\pm 1.58}$	$95.01_{\pm 4.40}$	$54.20_{\pm 1.19}$
	FeatTrans	$38.45_{\pm 0.42}$	$98.05_{\pm 1.04}$	$38.45_{\pm 0.43}$	$38.20_{\pm 0.86}$	$97.06_{\pm 1.39}$	$54.81_{\pm 0.72}$
	IACS	$72.10_{\pm 5.43}$	$85.44_{\pm 4.30}$	$72.10_{\pm 1.58}$	$66.29_{\pm 3.99}$	$84.90_{\pm 0.58}$	$73.68_{\pm 2.88}$
	IACS-G	$67.76_{\pm 3.00}$	$86.17_{\pm 1.65}$	$67.76_{\pm 1.80}$	$64.86_{\pm 1.81}$	$86.36_{\pm 3.48}$	$74.05_{\pm 1.94}$
	IACS-P	$72.38_{\pm 4.80}$	$83.69_{\pm 6.53}$	$72.38_{\pm 1.75}$	$65.14_{\pm 0.85}$	$86.07_{\pm 2.36}$	$74.28_{\pm 1.19}$

ACS

a) In general, IACS achieves the highest F1 score in most cases (5 out of 6), even when the graphs of training and inference are from different datasets.

b) The improvement in the 8-shot setting is relatively lower compared to the 4-shot setting.

Experiential Studies: Efficiency

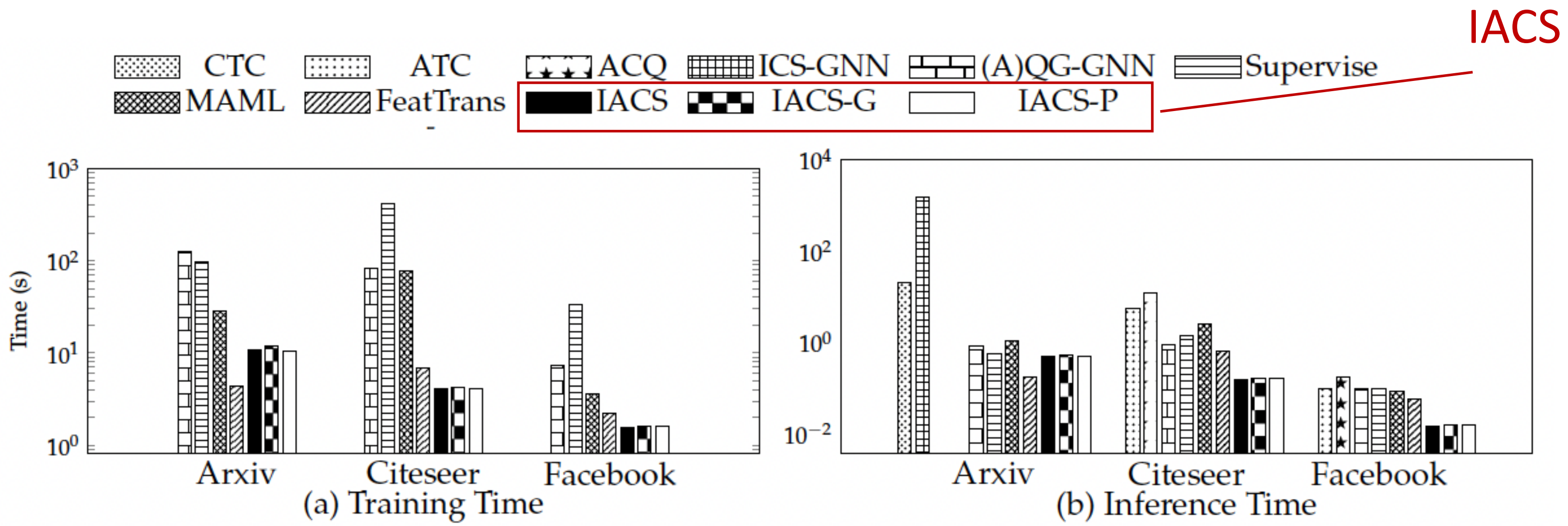


Figure 5: Comparison of Training and Inference Time

- IACS models exhibit faster training and inference time compared to other baselines in most datasets.

Experiential Studies: Streaming Model Adaptation

- The streaming adaptation process leads to higher F1 scores for the streaming model than the original model across a wide range of sequential tasks.
- The results indicate that three IACS models exhibit an improvement ratio of 3% in the streaming adaptation model.

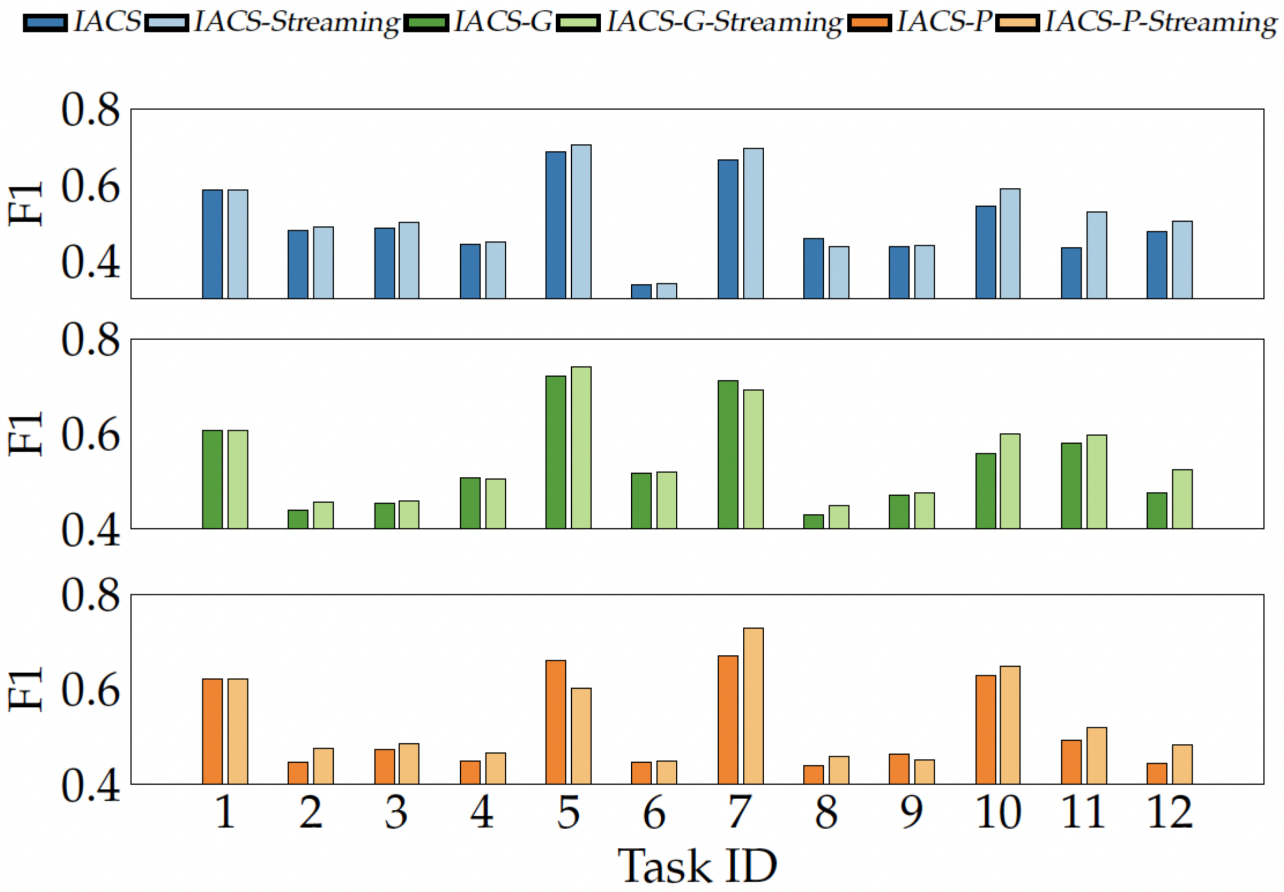
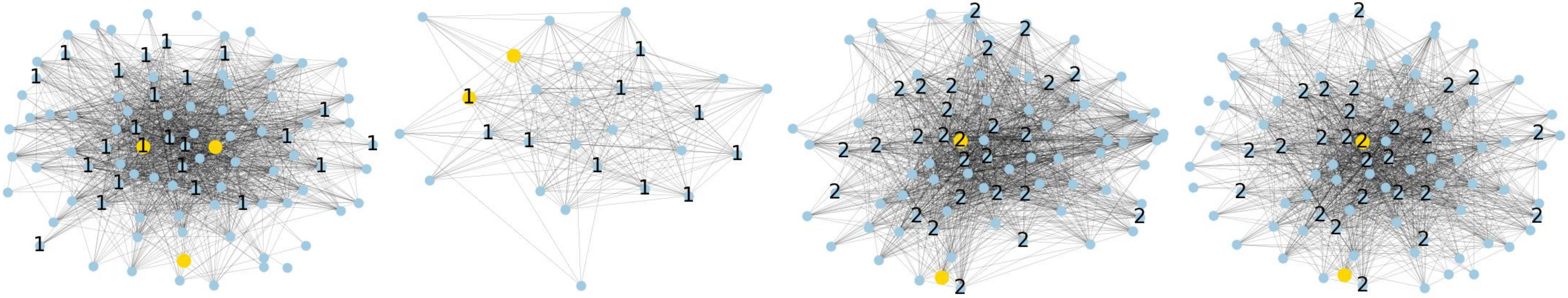


Figure 8: Streaming Model Adaptation on Twitter

Experiential Studies: Case Study



(a) Twitter: Training Communities (b) Ground-truth Community (c) Predicted Community

- These communities exhibit variations, characterized by heterogeneous topological structures and attribute distributions.
- We observe a notable overlap between the identified communities and the ground-truth, thus confirming the accuracy of our predictions.

Summary

- Leveraging ML/DL based approaches for attributed community search
- Existing learning approaches have limited inductive ability and cannot deal with complex queries.
- Propose Inductive Attributed Community Search (IACS) to infer new queries for different communities/graphs
- Propose a three-stage workflow to fulfill inductive ACS
- IACS achieves better performance on effectiveness and efficiency

Thank you!

Shuheng Fang, The Chinese University of Hong Kong.

shfang@se.cuhk.edu.hk 😃