Santiago
Chile

SIGMOD
PODS
2024

SIGMOD
PODS

Association
for Computing
Machinery

Neural Attributed Community Search at Billion Scale

Jianwei Wang, Kai Wang, Xuemin Lin, Wenjie Zhang, Ying Zhang

jianwei.wang1@unsw.edu.au, w.kai@sjtu.edu.cn, xuemin.lin@sjtu.edu.cn,
wenjie.zhang@unsw.edu.au, ying.zhang@uts.edu.au

UNSW SYDNEY
上海交通大學 SHANGHAI JIAO TONG UNIVERSITY
UTS

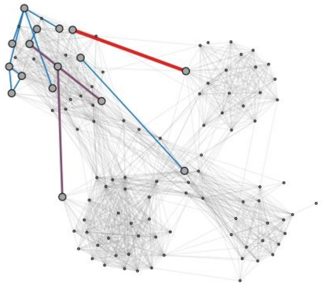
Outline

- Background and Problem Definition
- Motivations
- Methods
- Experiments
- Summary

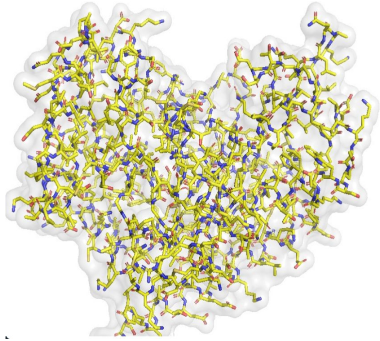
Background

- **Graph is everywhere.**

Cora (citation graph) [1]



Social graph [2]



Protein graph [3]

✔ Nodes are often featured with attributes

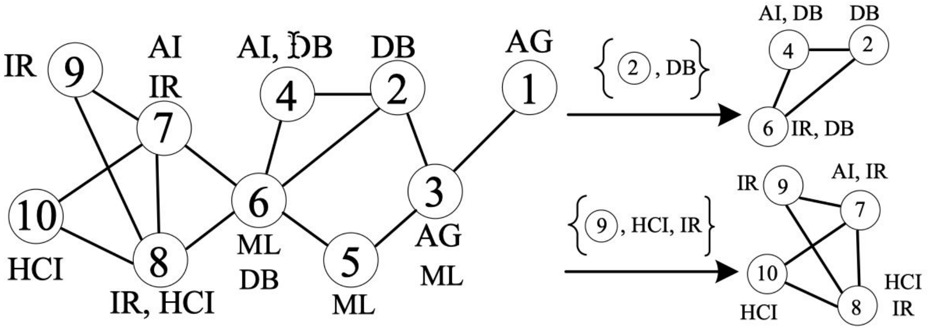
- **Community:** Normally, a set of nodes that are densely connected internally and loosely connected externally.

[1]: https://arxiv.org/pdf/2305.18405.pdf [2]: https://arxiv.org/pdf/1401.7233.pdf [3]: https://arxiv.org/pdf/2302.12177.pdf

Problem definition

- **Attributed Community Search:** Given an attributed graph $G(V, E, F)$, and a query $q = \langle V_q, F_q \rangle$ where $V_q \subseteq V$ is a set of query nodes and $F_q = \subseteq F$ is a set of query attributes, the task of attributed community search (ACS) aims to find a query-dependent community which preserves both structure cohesiveness and attribute homogeneity.

- **Applications**
  - Research communities mining.
  - Friend recommendation.
  - Protein complex identification.



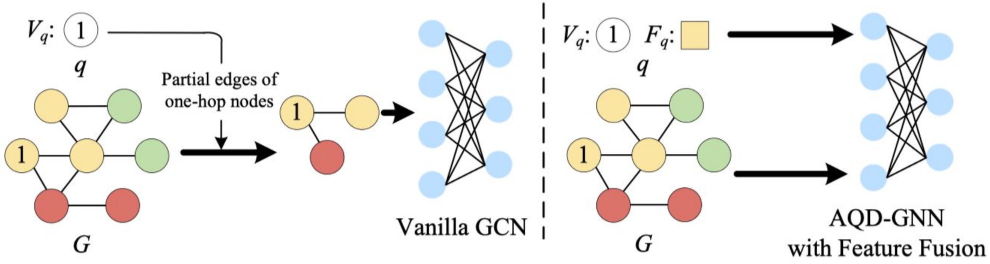
Motivation

- **Existing non-learning methods:**
  - ➢*k*-core based ACS model
  - ➢*k*-truss based ACS model

Structure Inflexibility

Attribute Irrelevance

- **Existing learning-based methods:**



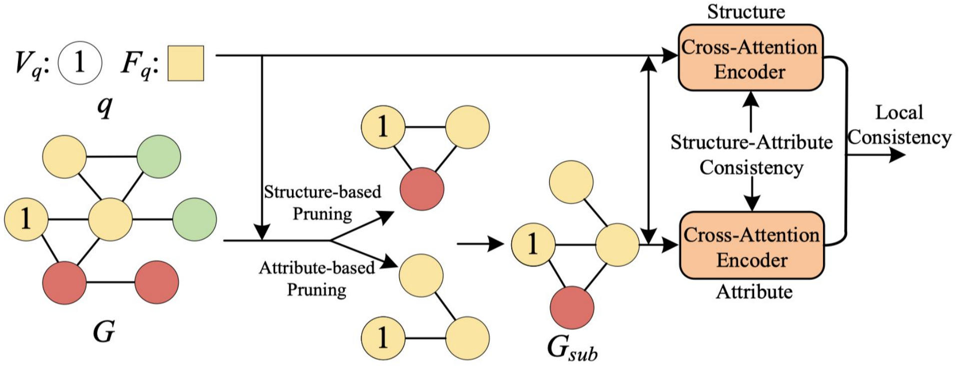
(a) ICS-GNN (one iteration)

(b) AQD-GNN

Efficiency and scalability issue for AQD-GNN

Interdependence among entities

Our methods



- **Candidate Subgraph Extraction**
  - ✓ Structure-based pruning with density sketch modularity
  - ✓ Attribute-based pruning
- **Consistency-aware Net (CoNet):**
  - ✓ Cross-Attention Encoder
  - ✓ Structure-Attribute Consistency & Local Consistency

Density sketch modularity

- Graph Modularity is a widely used measure for community cohesiveness. A higher modularity indicates a more cohesive community
  - Classical Modularity $CM(G,C) = \frac{1}{2|E|}\left(2|E_C| - \frac{d_C^2}{2|E|}\right)$
  - Density Modularity $DM(G,C) = \frac{1}{2|V_C|}\left(2|E_C| - \frac{d_C^2}{2|E|}\right)$
  - Density Sketch Modularity $DSM(G,C) = \frac{1}{2|V_C|^{\tau}}\left(2|E_C| - \frac{d_C^2}{2|E|}\right)$

sum of node degree

\# of internal nodes

τ is a hyperparameter to control granularity

\# of internal edges

\# of edges

- It checks the difference between the number of internal edges in the community and the number of expected edges in the community
- When τ approximates zero, density sketch modularity is as power as classical modularity
- When τ approximates one, density sketch modularity is as power as density modularity

Analysis of density sketch modularity

✓ When employing classic modularity for CS , it suffers from the free-rider effect and the resolution limit problem

- **Free-rider effect**

Given a set of query q, let C be a community identified based on a goodness function f, and C^* be the optimal solution (either local or global). The goodness function is said to be affected by the free-rider effect if $f(C \cup C^*) \geq f(C)$.

✓ Resulting community may encompass numerous nodes unrelated to the query nodes

- **Resolution limit problem**

Given a graph G, query q, the objective function f, a community constraint T, a community C satisfying T and containing all the query q, and any community C' satisfying the constraint T such that $C \cup C'$ is connected and $C \cap C' = \varnothing$, the objective function is said to suffer from the resolution limit problem if there exists a community C' such that $C \cup C'$ satisfies the constraint T and $f(C \cup C') \geq f(C)$.

✓ Resultant community may be too large to highlight some important structures.

Analysis of density sketch modularity

- For any positive τ, whenever density sketch modularity suffers from the free-rider effect, classic modularity suffers from the free-rider effect as well.

$$DSM(G, C \cup C*) \geq DSM(G, C)$$

$$\Rightarrow \frac{1}{2|V_{C\cup C^*}|^{\tau}}(2|E_{C\cup C^*}| - \frac{d^2_{C\cup C^*}}{2|E|}) \geq \frac{1}{2|V_C|^{\tau}}(2|E_C| - \frac{d^2_C}{2|E|})$$

As $2|V_C|^{\tau} > 0$

$$\Rightarrow \{\frac{|V_C|}{|V_{C\cup C^*}|}\}^{\tau}(2|E_{C\cup C^*}| - \frac{d^2_{C\cup C^*}}{2|E|}) \geq 2|E_C| - \frac{d^2_C}{2|E|}$$

$$\Rightarrow 2|E_{C\cup C^*}| - \frac{d^2_{C\cup C^*}}{2|E|} \geq \{\frac{|V_C|}{|V_{C\cup C^*}|}\}^{\tau}(2|E_{C\cup C^*}| - \frac{d^2_{C\cup C^*}}{2|E|}) \geq 2|E_C| - \frac{d^2_C}{2|E|}$$

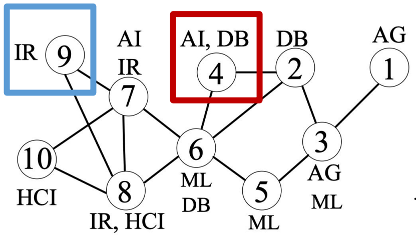
$$\Rightarrow CM(G, C \cup C^*) \geq CM(G, C)$$

- For any positive τ, whenever density sketch modularity suffers from the resolution-limit problem, classic modularity suffers from the resolution-limit problem as well.

Candidate subgraph extraction

- **Structure-based pruning**

  ✓ *k*-hop neighborhood with largest density sketch modularity (adaptively)



➢1-hop DSM: 0.504

➢2-hop DSM: 0.507

➢3-hop DSM: 0.135

➢1-hop DSM: 0.504

➢2-hop DSM: −0.094

➢3-hop DSM: 0.0

- **Attribute-based pruning:**

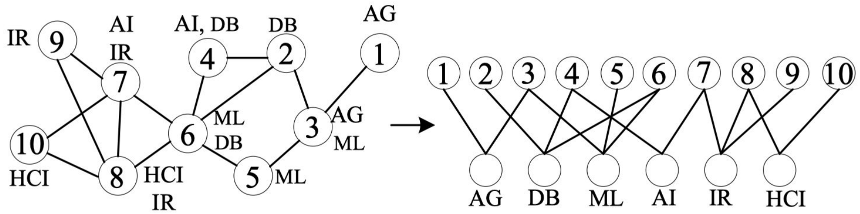


Figure 4: node-attribute bipartite graph

✓ *k*-hop neighborhood with largest bipartite modularity in the node-attribute bipartite graph

CoNet architecture

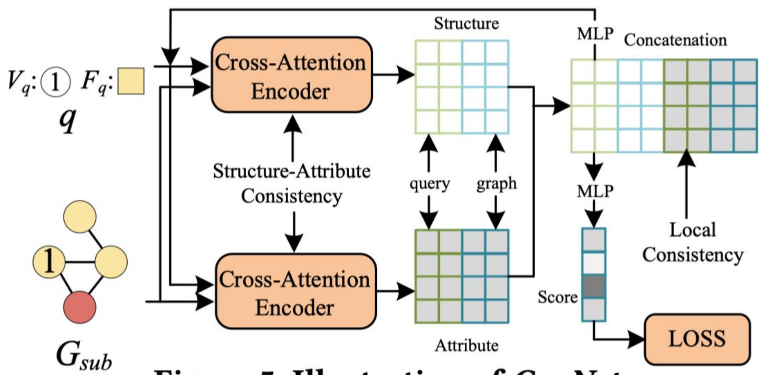


Figure 5: Illustration of *ConNet*

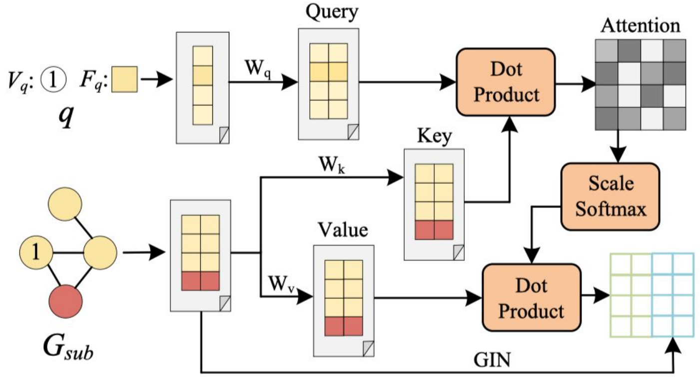


Figure 6: Illustration of Cross Attention Encoder

- Query Encoding $X_q = H_{v_q}^{(k)} W_q^{(s,k)},\ X_k = H^{(s,k)} W_k^{(s,k)},\ X_v = H^{(s,k)} W_v^{(s,k)}$

$$X = \mathrm{softmax}\left(\frac{X_q X_k^T}{\sqrt{d_{k+1}}}\right),\ H_{v_q}^{(k+1)} = X X_v$$

- Graph Encoding $h_v^{(s,k+1)} = \mathrm{MLP}^{(s,k)}\left(\left(1+\epsilon^{(k)}\right)\cdot h_v^{(s,k)},\ \sum_{v'\in N(v)} h_{v'}^{(s,k)}\right)$

- Lemma: ConNet is as powerful as the 1-WL algorithm.

Training Objectives

- **Structure-Attribute Consistency**

  ✓ Minimize the Wasserstein-1 distance between structure distribution and attribute distribution

$$W_1(\mathbb{P}_s, \mathbb{P}_a) = \inf_{\gamma \in \pi(\mathbb{P}_s, \mathbb{P}_a)} \mathbb{E}_{(\mu,\nu)\sim\gamma}[||\mu - \nu||]$$

$$W_1(\mathbb{P}_s, \mathbb{P}_a) = \sup_{||f_w||_L \leq 1} \mathbb{E}_{\mu\sim\mathbb{P}_s}[f_w(\mu)] - \mathbb{E}_{\nu\sim\mathbb{P}_a}[f_w(\nu)]$$

$$\mathcal{L}_w(H^{(s)}, H^{(a)}) = \sum_{h_v^{(a)} \in H^{(a)}} f_w(h_v^{(a)}) - \sum_{h_u^{(s)} \in H^{(s)}} f_w(h_u^{(s)})$$

$$\mathcal{L} = \mathcal{L}_b + \alpha\mathcal{L}_w + \beta\mathcal{L}_m$$

- **Local Consistency**

  ✓ Neighboring nodes have similar prediction

$$\mathcal{L}_m(H, A) = \left|\left|A - HH^T\right|\right|_F$$

- **Ground-truth information**

$$\mathcal{L}_b(\tilde{C}_{q_1}, C_{q_i}) = \sum_{j=1}^{|V_{sub}|} -C_{q_i,j} log(\tilde{C}_{q_1,j}) + (1 - C_{q_i,j}) log(1 - \tilde{C}_{q_1,j})$$

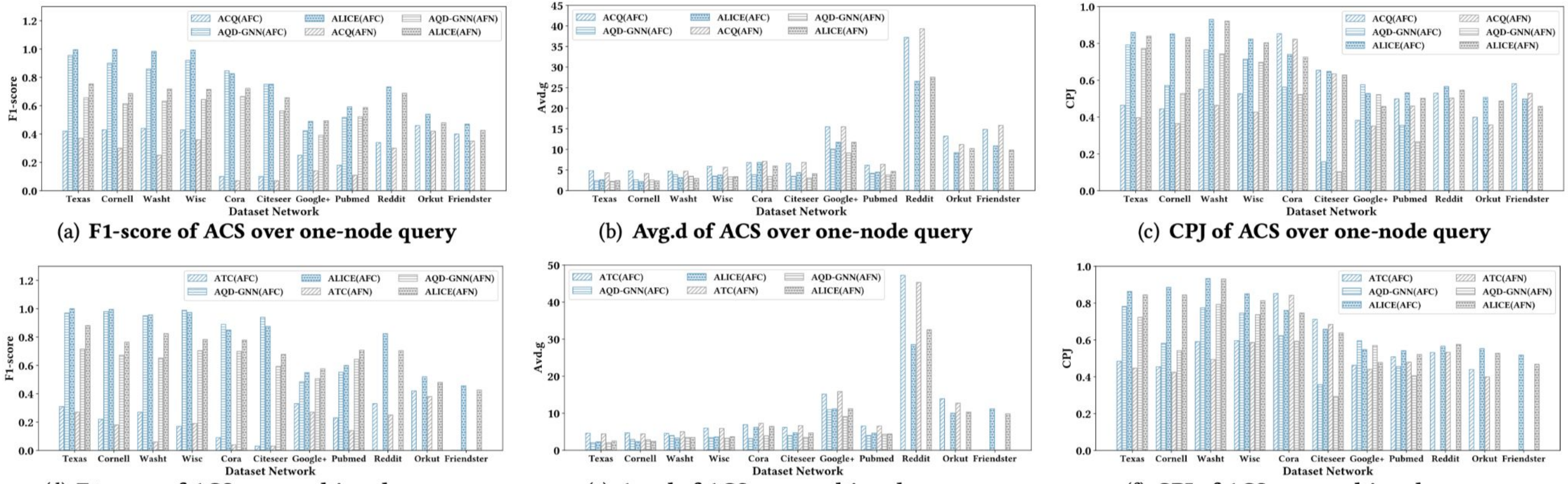
Experiments

Table 2: Statistics of the datasets

Dataset	$\|V\|$	$\|E\|$	$\|F^d\|$	N_c
Texas	187	279	1703	5
Cornell	195	285	1703	5
Washt	230	392	1703	5
Wiscs	265	469	1703	5
Cora	2708	5429	1433	7
Citeseer	3312	4715	3703	6
Google+	7856	321,268	2024	91
PubMed	19,717	44,324	500	3
Reddit	232,965	47,396,905	602	41
Orkut	3,072,627	117,185,083	1000	5000
Friendster	65,608,366	1,806,067,135	1000	5000

- **Query settings**
  - ✓ Attribute from communities (AFC)
  - ✓ Attribute from query node (AFN)
  - ✓ Empty attribute query (EmA)
- **Metrics**
  - ✓ F1-score
  - ✓ Average degree (Avg.d)
  - ✓ Community pair-wise Jaccard (CPJ)

Experiments



(a) F1-score of ACS over one-node query

(b) Avg.d of ACS over one-node query

(c) CPJ of ACS over one-node query

(d) F1-score of ACS over multi-node query

(e) Avg.d of ACS over multi-node query

(f) CPJ of ACS over multi-node query

Figure 7: Result on attributed community search

- Learning-based method has an average improvement of 54.50% in F1-score compared with traditional ACS method.
- ALICE has an average improvement of 10.18% compared to SOTA AQD-GNN using AFN as the query attribute.

Experiments

Table 3: Efficiency evaluation on different datasets (in seconds)

Method	*Texas*	*Cornell*	*Washt*	*Wisc*	*Cora*	*Citeseer*	*Google+*	*Pubmed*	*Reddit*	*Orkut*	*Frienster*
ICS-GNN (Train)	***	***	***	***	***	***	***	***	***	***	***
AQD-GNN (Train)	2.2+233	2.1+234	2.5+239	2.9+232	64.1+2214	59.3+4390	834.6+10035	3171.8+37059	—	—	—
ALICE (Train)	2.6+344	2.5+381	3.8+332	1.8+324	16.32+509	59.8+1239	189.8+3256	123.5+4317	8681+1107	2594.8+2224	65415.6+1244
ICS-GNN (Query)	20.5	25.1	27.4	28.6	167.7	124.3	627.6	112.3	1034.7	1540.8	24253.7
AQD-GNN (Query)	0.015+0.0021	0.014+0.0020	0.017+0.0022	0.019+0.0020	0.427+0.0026	0.395+0.0019	5.564+0.0019	21.14+0.0019	—	—	—
ALICE (Query)	0.017+0.0053	0.017+0.0045	0.025 + 0.0044	0.014 +0.0050	0.104+0.0041	0.398+0.0047	1.26+0.0053	0.823+0.0058	5.78+0.0052	17.29+0.0045	436.1+0.0048

(1) : We report preparation time + train (query) time; (2) : — indicates out of memory or not finished within 7 days; (3) : *** indicates this cell not applicable to this model.

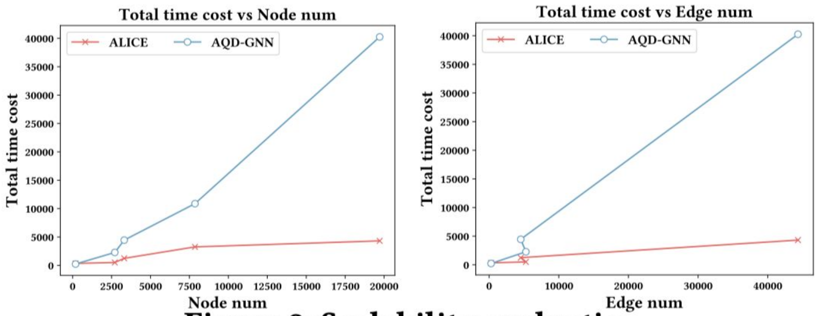


Figure 9: Scalability evaluation

- ✔ ALICE can deal with billion-scale graph while AQD-GNN cannot
- ✔ ALICE has a better scalability.

Experiments

Lw and Lb has an average improvement of 2.71% and 2.16% under complete labels, 4.57% and 4.03% under incomplete labels, respectively

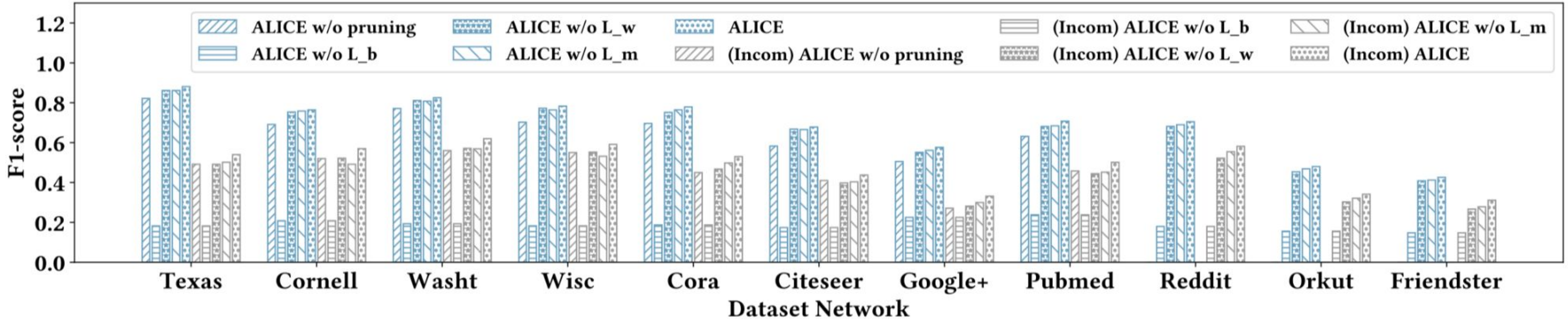
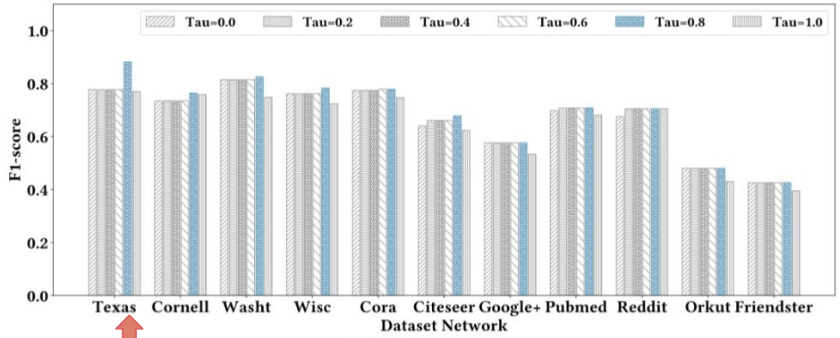
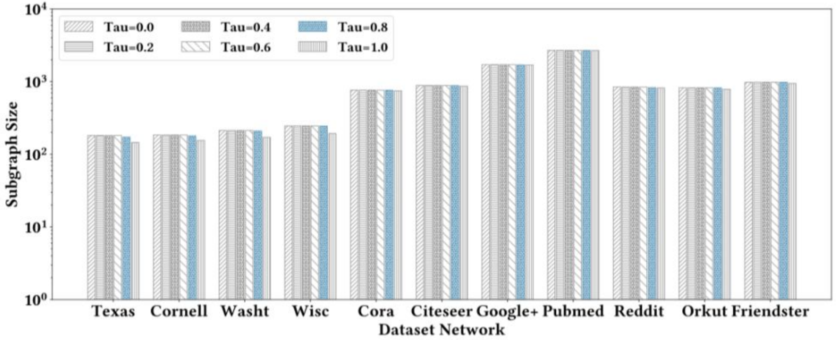


Figure 10: Ablation study



(a) **Accuracy**



(b) **Subgraph size**

Figure 11: Comparison of different modularity

Tau=0.8 has the best performance

Summary

- **We propose a learning-based framework, named ALICE, for attributed community search at large scale.**

- **We design an efficient subgraph extraction algorithm by leveraging density sketch modularity and node-attribute relationship to adaptively select promising nodes.**

- **We propose a GNN-based model ConNet to preserve both structure-attribute consistency and local consistency among nodes.**

- **Extensive experiments over 11 popular public datasets, encompassing one billion-scale graph Friendster, demonstrate the effectiveness of ALICE.**

Q & A