VERSACHI: Finding Statistically Significant Subgraph Matches using Chebyshev's Inequality

SHUBHANGI AGARWAL[†]
sagarwal@cse.iitk.ac.in

SOURAV DUTTA[††]
sourav.dutta2@huawei.com

ARNAB BHATTACHARYA[†]
arnabb@cse.iitk.ac.in

November, 2021

CIKM 2021, Gold Coast, Queensland, Australia (online)

[†]Department of Computer Science and Engineering,
Indian Institute of Technology Kanpur, **India**

[††]Huawei Research Centre, Dublin, **Ireland**

Introduction

Subgraph querying is useful in frequent pattern mining, community detection, question answering etc.

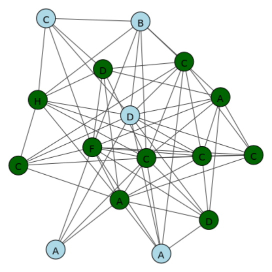


FIGURE 1: GRAPH (WITH QUERY MANIFESTATION)

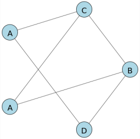


FIGURE 2: QUERY

Introduction

Subgraph querying is useful in frequent pattern mining, community detection, question answering etc.

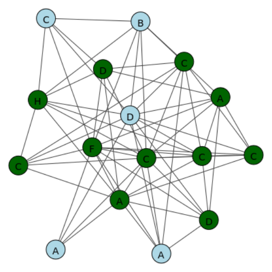


FIGURE 1: GRAPH (WITH QUERY MANIFESTATION)

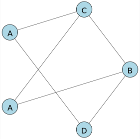


FIGURE 2: QUERY

- Two types of matches:
  - Exact matches
  - Approximate Subgraph Matching

Introduction

Subgraph querying is useful in frequent pattern mining, community detection, question answering etc.

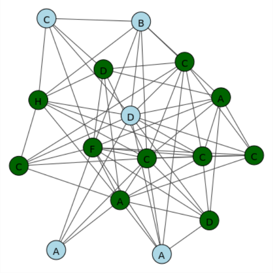


FIGURE 1: GRAPH (WITH QUERY MANIFESTATION)

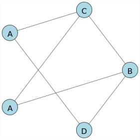


FIGURE 2: QUERY

- Two types of matches:
  - Exact matches
  - **Approximate Subgraph Matching**

Introduction

Subgraph querying is useful in frequent pattern mining, community detection, question answering etc.

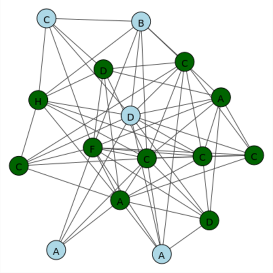


FIGURE 1: GRAPH (WITH QUERY MANIFESTATION)

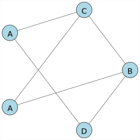


FIGURE 2: QUERY

- Two types of matches:
  - Exact matches
  - **Approximate Subgraph Matching**
- Similarity metric?

Introduction

Subgraph querying is useful in frequent pattern mining, community detection, question answering etc.

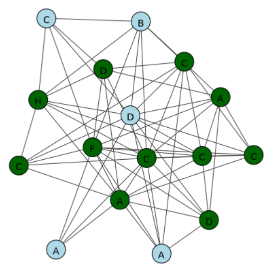


FIGURE 1: GRAPH (WITH QUERY MANIFESTATION)

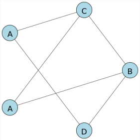


FIGURE 2: QUERY

- Two types of matches:
  - Exact matches
  - **Approximate Subgraph Matching**
- Similarity metric?
  - Statistical Significance (χ^2)
- Capture underlying distribution
  - Chebyshev's Inequality

Aim

Goal: Find top-k best approximate matches of the query in the target graph.

Aim

Goal: Find top-k best approximate matches of the query in the target graph.

VerSaChI - *Vertex Neighborhood Aggregation for Statistically Significant Subgraphs via Chebyshev's Inequality*

Aim

Goal: Find top-k best approximate matches of the query in the target graph.

VerSaChI - *Vertex Neighborhood Aggregation for Statistically Significant Subgraphs via Chebyshev's Inequality*

- Capture two-hop similarity of $v \in \mathcal{G}$ with $q \in \mathcal{Q}$

Aim

Goal: Find top-k best approximate matches of the query in the target graph.

VERSACHI - *Vertex Neighborhood Aggregation for Statistically Significant Subgraphs via Chebyshev's Inequality*

- Capture two-hop similarity of $v \in \mathcal{G}$ with $q \in \mathcal{Q}$
- Chebyshev's Inequality
  - Probability of deviation based on mean and standard deviation

Aim

Goal: Find top-k best approximate matches of the query in the target graph.

VERSACHI - *Vertex Neighborhood Aggregation for Statistically Significant Subgraphs via Chebyshev's Inequality*

- Capture two-hop similarity of $v \in \mathcal{G}$ with $q \in \mathcal{Q}$
- Chebyshev's Inequality
  – Probability of deviation based on mean and standard deviation
- Pearson's χ^2 statistic
  – Deviation of observed from expected underlying distribution

VERSACHI

Offline Phase

1. Create indexes
2. Neighbor similarity ($\eta_{u,v}$)
3. Graph characteristics
4. Discretize values (into category symbols, σ_i)
5. Probability of σ_i (using Chebyshev)

Online (Querying) Phase

i. Create candidate pairs $\langle v, q \rangle$ (using indexes)

ii. Symbol sequence (1^{st} and 2^{nd} hop neighbors)

iii. Compute similarity of $\langle v, q \rangle$ (statistical significance)

iv. Greedy expansion

VERSACHI

Offline Phase

1. Create indexes
2. Neighbor similarity ($\eta_{u,v}$)
3. Graph characteristics
4. Discretize values
   $\big($into category symbols, $\sigma_i\big)$
5. Probability of σ_i
   $\big($using Chebyshev$\big)$

Online (Querying) Phase

i. Create candidate pairs $\langle v, q\rangle$
   $\big($using indexes$\big)$

ii. Symbol sequence
   $\big(1^{st}$ and 2^{nd} hop neighbors$\big)$

iii. Compute similarity of $\langle v, q\rangle$
   $\big($statistical significance$\big)$

iv. Greedy expansion

VERSACHI

Offline Phase

1. Create indexes
2. Neighbor similarity ($\eta_{u,v}$)
3. Graph characteristics
4. Discretize values $\big($into category symbols, $\sigma_i\big)$
5. Probability of σ_i $\big($using Chebyshev$\big)$

Online (Querying) Phase

i. Create candidate pairs $\langle v, q\rangle$ $\big($using indexes$\big)$

ii. Symbol sequence $\big(1^{st}$ and 2^{nd} hop neighbors$\big)$

iii. Compute similarity of $\langle v, q\rangle$ $\big($statistical significance$\big)$

iv. Greedy expansion

Offline Phase

Steps

1. Create indexes
2. Neighbor similarity
3. Graph characteristics
4. Discretize values
5. Probability of σ_i

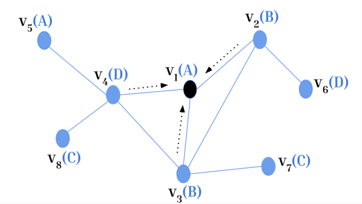


FIGURE 3: EXAMPLE TARGET GRAPH ($\mathcal{G}$)

Offline Phase

Steps

1. Create indexes
2. Neighbor similarity
3. Graph characteristics
4. Discretize values
5. Probability of σ_i

1. Inverted Label index ($IL_{\mathcal{G}}$)
   – map labels to vertices

2. Neighbor Label index ($NL_{\mathcal{G}}$)
   – labels of neighbors

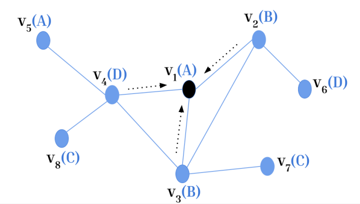


FIGURE 3: EXAMPLE TARGET GRAPH ($\mathcal{G}$)

$IL_{\mathcal{G}}$ index

Label	Vertex
A	$\{v_1, v_5\}$
B	$\{v_2, v_3\}$
C	$\{v_7, v_8\}$
D	$\{v_4, v_6\}$

– Neighbors of $v_1 = \{v_2, v_3, v_4\}$

– $NL_{\mathcal{G}}$ index of v_1

$$\mathcal{N}(v_1) = \{A, B, B, D\}$$

Offline Phase

Steps

1. Create indexes
2. Neighbor similarity
3. Graph characteristics
4. Discretize values
5. Probability of σ_i

1. Capture neighborhood overlap – Penalize absence of neighbor labels
2. Done $\forall u, v \in \mathcal{G}$
3. Modified *Tversky index* ($\gamma = 3$)

$$\eta_{u,v} = \frac{|\mathcal{N}(u) \cap \mathcal{N}(v)|}{|\mathcal{N}(u) \cap \mathcal{N}(v)| + |\mathcal{N}(v) \setminus \mathcal{N}(u)|^{\gamma}}$$

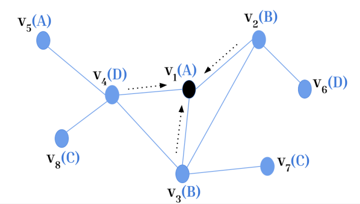


FIGURE 3: EXAMPLE TARGET GRAPH ($\mathcal{G}$)

$$\eta_{\mathrm{v}_1,\mathrm{v}_4} = \frac{|\mathcal{N}(\mathrm{v}_1) \cap \mathcal{N}(\mathrm{v}_4)|}{|\mathcal{N}(\mathrm{v}_1) \cap \mathcal{N}(\mathrm{v}_4)| + |\mathcal{N}(\mathrm{v}_4) \setminus \mathcal{N}(\mathrm{v}_1)|^3}$$

$$= \frac{3}{(3 + 2^3)} = \frac{3}{11}$$

$$\mathcal{N}(\mathrm{v}_1) = \{\mathrm{A, B, B, D}\}$$

$$\mathcal{N}(\mathrm{v}_4) = \{\mathrm{D, A, B, C, A}\}$$

Offline Phase

Steps

1. Create indexes
2. Neighbor similarity
3. Graph characteristics
4. Discretize values
5. Probability of σ_i

1. $\psi(\mathcal{G})$: Avg. neighbor similarity

$$\psi(\mathcal{G}) = 0.59$$

2. $\delta(\mathcal{G})$: Std. Dev. of similarity

$$\delta(\mathcal{G}) = 0.38$$

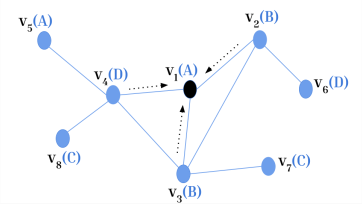


FIGURE 3: EXAMPLE TARGET GRAPH ($\mathcal{G}$)

3. $\Delta(\mathcal{G})$: Maximum z-score

$$\Delta(\mathcal{G}) = \max_{u,v \in \mathcal{G}} \left\{ \frac{|\eta_{u,v} - \psi(\mathcal{G})|}{\delta(\mathcal{G})} \right\}$$

$$\Delta(\mathcal{G}) = 1.54$$

Offline Phase

Steps

1. Create indexes
2. Neighbor similarity
3. Graph characteristics
4. Discretize values
5. Probability of σ_i

1. $\psi(\mathcal{G})$: Avg. neighbor similarity

$$\psi(\mathcal{G}) = 0.59$$

2. $\delta(\mathcal{G})$: Std. Dev. of similarity

$$\delta(\mathcal{G}) = 0.38$$

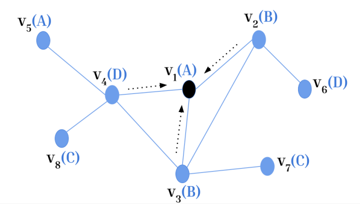


FIGURE 3: EXAMPLE TARGET GRAPH ($\mathcal{G}$)

3. $\Delta(\mathcal{G})$: Maximum z-score

$$\Delta(\mathcal{G}) = \max_{u,v \in \mathcal{G}} \left\{ \frac{|\eta_{u,v} - \psi(\mathcal{G})|}{\delta(\mathcal{G})} \right\}$$

$$\Delta(\mathcal{G}) = 1.54$$

Offline Phase

Steps

1. Create indexes
2. Neighbor similarity
3. Graph characteristics
4. Discretize values
5. Probability of σ_i

1. Capture degree of matching of a vertex pair
   – based on deviation from $\psi(\mathcal{G})$
2. Step size (κ)
   – Lower values preferred
3. # categories/ symbols:
   $\tau = \lceil (\Delta(\mathcal{G}) - 1)/\kappa \rceil$

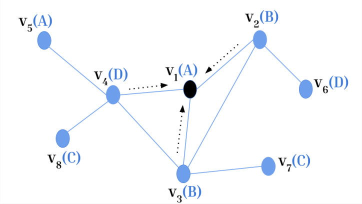


FIGURE 3: EXAMPLE TARGET GRAPH ($\mathcal{G}$)

– Symbol set = $\{\sigma_1, \sigma_2, \ldots, \sigma_\tau\}$

$$\sigma_1 : 0 \leq \frac{|\eta_{u,v} - \psi(\mathcal{G})|}{\delta(\mathcal{G})} < 1 + \kappa$$

$$\sigma_{i \in [2,\tau]} : 1 + (i-1) \cdot \kappa \leq \frac{|\eta_{u,v} - \psi(\mathcal{G})|}{\delta(\mathcal{G})} < 1 + i \cdot \kappa$$

For $\kappa = 0.1$,

$$\tau = \left\lceil \frac{(1.54 - 1)}{0.1} \right\rceil = \lceil 5.4 \rceil = 6$$

$$\frac{\eta_{v_1, v_4} - \psi(\mathcal{G})}{\delta(G)} = \frac{\frac{3}{11} - 0.59}{0.38} = -0.83$$

Symbol assigned: σ_1

Offline Phase

Steps

1. Create indexes
2. Neighbor similarity
3. Graph characteristics
4. Discretize values
5. Probability of σ_i

1. Capture degree of matching of a vertex pair
   – based on deviation from $\psi(\mathcal{G})$
2. Step size (κ)
   – Lower values preferred
3. # categories/ symbols:
   $\tau = \lceil (\Delta(\mathcal{G}) - 1)/\kappa \rceil$

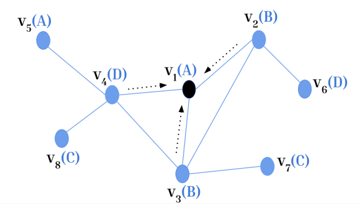


FIGURE 3: EXAMPLE TARGET GRAPH ($\mathcal{G}$)

– Symbol set = $\{\sigma_1, \sigma_2, \ldots, \sigma_\tau\}$

$$\sigma_1 : 0 \leq \frac{|\eta_{u,v} - \psi(\mathcal{G})|}{\delta(\mathcal{G})} < 1 + \kappa$$

$$\sigma_{i \in [2,\tau]} : 1 + (i-1) \cdot \kappa \leq \frac{|\eta_{u,v} - \psi(\mathcal{G})|}{\delta(\mathcal{G})} < 1 + i \cdot \kappa$$

For $\kappa = 0.1$,

$$\tau = \left\lceil \frac{(1.54-1)}{0.1} \right\rceil = \lceil 5.4 \rceil = 6$$

$$\frac{\eta_{v_1,v_4} - \psi(\mathcal{G})}{\delta(G)} = \frac{\frac{3}{11} - 0.59}{0.38} = -0.83$$

Symbol assigned: σ_1

Offline Phase

Steps

1. Create indexes
2. Neighbor similarity
3. Graph characteristics
4. Discretize values
5. Probability of σ_i

1. Capture degree of matching of a vertex pair
   – based on deviation from $\psi(\mathcal{G})$
2. Step size (κ)
   – Lower values preferred
3. # categories/ symbols:
   $\tau = \lceil (\Delta(\mathcal{G}) - 1)/\kappa \rceil$

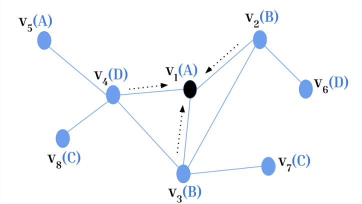


FIGURE 3: EXAMPLE TARGET GRAPH ($\mathcal{G}$)

– Symbol set = $\{\sigma_1, \sigma_2, \ldots, \sigma_\tau\}$

$$\sigma_1 : 0 \leq \frac{|\eta_{u,v} - \psi(\mathcal{G})|}{\delta(\mathcal{G})} < 1 + \kappa$$

$$\sigma_{i \in [2,\tau]} : 1 + (i-1) \cdot \kappa \leq \frac{|\eta_{u,v} - \psi(\mathcal{G})|}{\delta(\mathcal{G})} < 1 + i \cdot \kappa$$

For $\kappa = 0.1$,

$$\tau = \left\lceil \frac{(1.54 - 1)}{0.1} \right\rceil = \lceil 5.4 \rceil = 6$$

$$\frac{\eta_{v_1,v_4} - \psi(\mathcal{G})}{\delta(G)} = \frac{\frac{3}{11} - 0.59}{0.38} = -0.83$$

Symbol assigned: σ_1

Offline Phase

Steps

1. Create indexes
2. Neighbor similarity
3. Graph characteristics
4. Discretize values
5. Probability of σ_i

1. Capture degree of matching of a vertex pair
   – based on deviation from $\psi(\mathcal{G})$
2. Step size (κ)
   – Lower values preferred
3. # categories/ symbols:
   $\tau = \lceil(\Delta(\mathcal{G}) - 1)/\kappa\rceil$

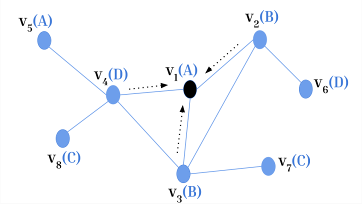


FIGURE 3: EXAMPLE TARGET GRAPH ($\mathcal{G}$)

– Symbol set = $\{\sigma_1, \sigma_2, \ldots, \sigma_\tau\}$

$$\sigma_1 : 0 \leq \frac{|\eta_{u,v} - \psi(\mathcal{G})|}{\delta(\mathcal{G})} < 1 + \kappa$$

$$\sigma_{i\in[2,\tau]} : 1 + (i-1)\cdot\kappa \leq \frac{|\eta_{u,v} - \psi(\mathcal{G})|}{\delta(\mathcal{G})} < 1 + i\cdot\kappa$$

For $\kappa = 0.1$,

$$\tau = \left\lceil \frac{(1.54-1)}{0.1} \right\rceil = \lceil 5.4 \rceil = 6$$

$$\frac{\eta_{v_1,v_4} - \psi(\mathcal{G})}{\delta(G)} = \frac{\frac{3}{11} - 0.59}{0.38} = -0.83$$

Symbol assigned: σ_1

Offline Phase

Steps

1. Create indexes
2. Neighbor similarity
3. Graph characteristics
4. Discretize values
5. Probability of σ_i

1. $\Pr(\sigma_i)$: Probability of symbol occurrence

2. Using Chebyshev's Inequality – $\Pr(\frac{|X-\mu|}{\delta} \geq t) \leq 1/t^2, \quad t > 0$

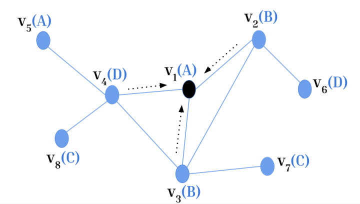


FIGURE 3: EXAMPLE TARGET GRAPH ($\mathcal{G}$)

$$\Pr(\sigma_i) = \frac{1}{2}\left[\frac{1}{(1+(i-1)\cdot\kappa)^2} - \frac{1}{(1+i\cdot\kappa)^2}\right], \qquad 2 \leq i \leq \tau$$

$$\Pr(\sigma_1) = 1 - \sum_{j=2}^{\tau} \Pr(\sigma_j)$$

$$\sum_{j=2}^{\tau=6} \Pr(\sigma_j) = 0.22 \qquad \Pr(\sigma_1) = 1 - 0.22 = 0.78$$

Offline Phase

Steps

1. Create indexes
2. Neighbor similarity
3. Graph characteristics
4. Discretize values
5. Probability of σ_i

1. $\Pr(\sigma_i)$: Probability of symbol occurrence

2. Using Chebyshev's Inequality – $\Pr(\frac{|X-\mu|}{\delta} \geq t) \leq 1/t^2, \quad t > 0$

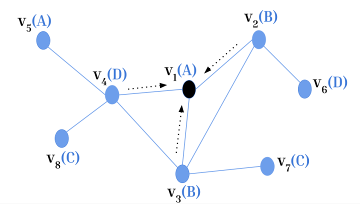


FIGURE 3: EXAMPLE TARGET GRAPH ($\mathcal{G}$)

$$\Pr(\sigma_i) = \frac{1}{2}\left[\frac{1}{(1+(i-1)\cdot\kappa)^2} - \frac{1}{(1+i\cdot\kappa)^2}\right], \qquad 2 \leq i \leq \tau$$

$$\Pr(\sigma_1) = 1 - \sum_{j=2}^{\tau} \Pr(\sigma_j)$$

$$\sum_{j=2}^{\tau=6} \Pr(\sigma_j) = 0.22 \qquad\qquad \Pr(\sigma_1) = 1 - 0.22 = 0.78$$

Querying Phase

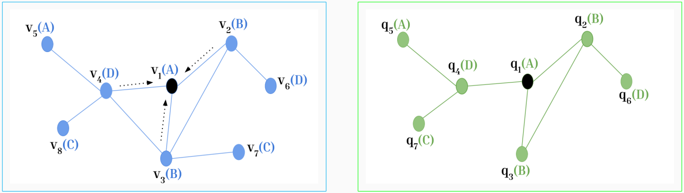
Steps

i. Create candidate pairs

ii. Symbol sequence

iii. Compute similarity of (u, q)

iv. Greedy expansion



(A) (B)

FIGURE 4: EXAMPLE (A) TRAGET GRAPH ($\mathcal{G}$) AND (B) QUERY GRAPH ($\mathcal{Q}$)

Querying Phase

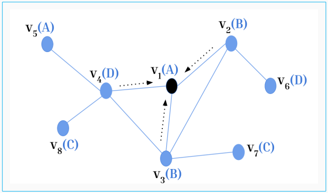
Steps

i. Create candidate pairs

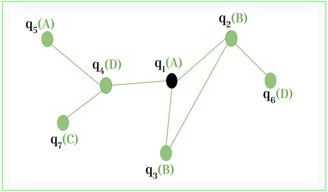
ii. Symbol sequence

iii. Compute similarity of (v, q)

iv. Greedy expansion



(A)



(B)

FIGURE 4: EXAMPLE (A) TRAGET GRAPH ($\mathcal{G}$) AND (B) QUERY GRAPH ($\mathcal{Q}$)

1. Create indexes $(IL_{\mathcal{Q}}, NL_{\mathcal{Q}})$ $\longrightarrow$ 2. Form Candidate Pairs

$IL_{\mathcal{Q}}$ index

Label	Vertex
A	$\{q_1, q_5\}$
B	$\{q_2, q_3\}$
C	$\{q_7\}$
D	$\{q_4, q_6\}$

– Neighbours of $q_1 = \{q_2, q_3, q_4\}$

– $NL_{\mathcal{Q}}$ index of q_1

$$\mathcal{N}(q_1) = \{A, B, B, D\}$$

Candidate vertex pairs $\langle v, q \rangle$

$v \in \mathcal{G}, q \in \mathcal{Q}$

A: $\langle v_1, q_1 \rangle, \langle v_1, q_5 \rangle,$
$\langle v_5, q_1 \rangle, \langle v_5, q_5 \rangle$

C: $\langle v_7, q_7 \rangle, \langle v_8, q_7 \rangle$

Querying Phase

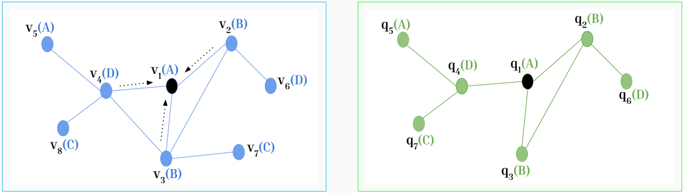
Steps

i. Create candidate pairs

ii. Symbol sequence

iii. Compute similarity of $\langle v, q \rangle$

iv. Greedy expansion



(A) (B)

FIGURE 4: EXAMPLE (A) TRAGET GRAPH ($\mathcal{G}$) AND (B) QUERY GRAPH ($\mathcal{Q}$)

1. Create indexes ($IL_{\mathcal{Q}}, NL_{\mathcal{Q}}$) $\longrightarrow$ 2. Form Candidate Pairs

$IL_{\mathcal{Q}}$ index

Label	Vertex
A	$\{q_1, q_5\}$
B	$\{q_2, q_3\}$
C	$\{q_7\}$
D	$\{q_4, q_6\}$

– Neighbours of $q_1 = \{q_2, q_3, q_4\}$

– $NL_{\mathcal{Q}}$ index of q_1

$$\mathcal{N}(q_1) = \{A, B, B, D\}$$

Candidate vertex pairs $\langle v, q \rangle$

$v \in \mathcal{G}, q \in \mathcal{Q}$

A: $\langle v_1, q_1 \rangle, \langle v_1, q_5 \rangle,$
$\langle v_5, q_1 \rangle, \langle v_5, q_5 \rangle$

C: $\langle v_7, q_7 \rangle, \langle v_8, q_7 \rangle$

Querying Phase

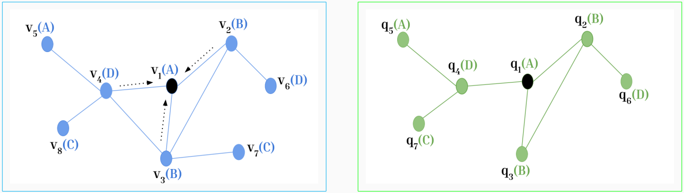
Steps

i. Create candidate pairs

ii. Symbol sequence

iii. Compute similarity of $\langle v, q \rangle$

iv. Greedy expansion



(A) (B)

FIGURE 4: EXAMPLE (A) TRAGET GRAPH ($\mathcal{G}$) AND (B) QUERY GRAPH ($\mathcal{Q}$)

For all vertex pairs $\langle$v, q$\rangle$

1. Compute $\eta_{\text{v, q}}$ $\longrightarrow$ 2. Compute 2^{nd} order neighbor similarity (Greedy best mapping based on η values) $\longrightarrow$ 3. Assign Symbol

$$\eta_{\mathrm{v}_1,\mathrm{q}_1} = 1$$

$$\langle \mathrm{v}_1, \mathrm{q}_1 \rangle : \sigma_1$$

Neighbour pair	Symbol
$\langle \mathrm{v}_2, \mathrm{q}_2 \rangle$	σ_2
$\langle \mathrm{v}_3, \mathrm{q}_3 \rangle$	σ_3
$\langle \mathrm{v}_4, \mathrm{q}_4 \rangle$	σ_4

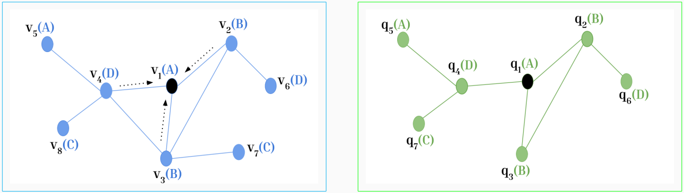
Vertex Symbol Sequence:

$$O_{\langle \mathrm{v}_1,\mathrm{q}_1 \rangle} = \{\sigma_1, \sigma_2, \sigma_3, \sigma_4\}$$

Querying Phase

Steps

i. Create candidate pairs
ii. Symbol sequence
iii. Compute similarity of $\langle v, q \rangle$
iv. Greedy expansion



(A) (B)

FIGURE 4: EXAMPLE (A) TRAGET GRAPH ($\mathcal{G}$) AND (B) QUERY GRAPH ($\mathcal{Q}$)

For all vertex pairs $\langle \text{v}, \text{q} \rangle$

1. Compute $\eta_{\text{v, q}}$ $\longrightarrow$ 2. Compute 2^{nd} order neighbor similarity (Greedy best mapping based on η values) $\longrightarrow$ 3. Assign Symbol

$\eta_{\text{v}_1, \text{q}_1} = 1$

$\langle \text{v}_1, \text{q}_1 \rangle : \sigma_1$

Neighbour pair	Symbol
$\langle \text{v}_2, \text{q}_2 \rangle$	σ_2
$\langle \text{v}_3, \text{q}_3 \rangle$	σ_3
$\langle \text{v}_4, \text{q}_4 \rangle$	σ_4

Vertex Symbol Sequence:

$O_{\langle \text{v}_1, \text{q}_1 \rangle} = \{\sigma_1, \sigma_2, \sigma_3, \sigma_4\}$

Querying Phase

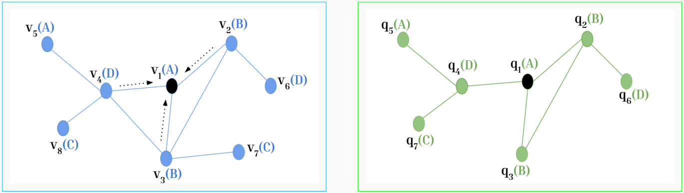
Steps

i. Create candidate pairs

ii. Symbol sequence

iii. Compute similarity of $\langle v, q \rangle$

iv. Greedy expansion



(A) (B)

FIGURE 4: EXAMPLE (A) TRAGET GRAPH ($\mathcal{G}$) AND (B) QUERY GRAPH ($\mathcal{Q}$)

For all vertex pairs $\langle \mathrm{v}, \mathrm{q} \rangle$

1. Compute $\eta_{\mathrm{v,\,q}}$ $\longrightarrow$ 2. Compute 2^{nd} order neighbor similarity (Greedy best mapping based on η values) $\longrightarrow$ 3. Assign Symbol

$\eta_{\mathrm{v}_1, \mathrm{q}_1} = 1$

$\langle \mathrm{v}_1, \mathrm{q}_1 \rangle : \sigma_1$

Neighbour pair	Symbol
$\langle \mathrm{v}_2, \mathrm{q}_2 \rangle$	σ_2
$\langle \mathrm{v}_3, \mathrm{q}_3 \rangle$	σ_3
$\langle \mathrm{v}_4, \mathrm{q}_4 \rangle$	σ_4

Vertex Symbol Sequence:

$O_{\langle \mathrm{v}_1, \mathrm{q}_1 \rangle} = \{\sigma_1, \sigma_2, \sigma_3, \sigma_4\}$

Querying Phase

Steps

i. Create candidate pairs
ii. Symbol sequence
iii. Compute similarity of $\langle v, q \rangle$
iv. Greedy expansion

1. Compute similarity $\left(\chi^2_{\langle \mathrm{v,\,q} \rangle}\right)$

$$\chi^2_{\langle \mathrm{v,\,q} \rangle} = \sum_{\forall i} \frac{\left[O_{\langle \mathrm{v,\,q} \rangle}(i) - E_{\langle \mathrm{v,\,q} \rangle}(i)\right]^2}{O_{\langle \mathrm{v,\,q} \rangle}(i)}$$

2. $O_{\langle \mathrm{v,\,q} \rangle}(i)$ - Observed count of σ_i; $E_{\langle \mathrm{v,\,q} \rangle}(i)$ - Expected count of σ_i

3. $O_{\langle \mathrm{v}_1, \mathrm{q}_1 \rangle} = \{\sigma_1, \sigma_2, \sigma_3, \sigma_4\}$; Length of $O_{\langle \mathrm{v}_1, \mathrm{q}_1 \rangle} = 3 + 1 = 4$

- $E_{\langle \mathrm{v,\,q} \rangle}(i) = l \cdot \Pr(\sigma_i)$
  - l = length of $O_{\langle \mathrm{v,\,q} \rangle}$

- $E_{\langle \mathrm{v}_1, \mathrm{q}_1 \rangle}(1) = 4 \cdot 0.78 = 3.12$
- $O_{\langle \mathrm{v}_1, \mathrm{q}_1 \rangle}(1) = 1$
- $\chi^2_{\langle \mathrm{v}_1, \mathrm{q}_1 \rangle}(1) = \frac{(1-3.12)^2}{3.12} = 1.44$

Querying Phase

Steps

i. Create candidate pairs
ii. Symbol sequence
iii. Compute similarity of $\langle v, q\rangle$
iv. Greedy expansion

1. Compute similarity $\left(\chi^2_{\langle \text{v, q}\rangle}\right)$

$$\chi^2_{\langle \text{v, q}\rangle} = \sum_{\forall i} \frac{\left[O_{\langle \text{v, q}\rangle}(i) - E_{\langle \text{v, q}\rangle}(i)\right]^2}{O_{\langle \text{v, q}\rangle}(i)}$$

2. $O_{\langle \text{v, q}\rangle}(i)$ - Observed count of σ_i; $E_{\langle \text{v, q}\rangle}(i)$ - Expected count of σ_i
3. $O_{\langle \text{v}_1, \text{q}_1\rangle} = \{\sigma_1, \sigma_2, \sigma_3, \sigma_4\}$; Length of $O_{\langle \text{v}_1, \text{q}_1\rangle} = 3 + 1 = 4$

- $E_{\langle \text{v, q}\rangle}(i) = l \cdot \Pr(\sigma_i)$
  - l = length of $O_{\langle \text{v, q}\rangle}$

- $E_{\langle \text{v}_1, \text{q}_1\rangle}(1) = 4 \cdot 0.78 = 3.12$
- $O_{\langle \text{v}_1, \text{q}_1\rangle}(1) = 1$
- $\chi^2_{\langle \text{v}_1, \text{q}_1\rangle}(1) = \frac{(1-3.12)^2}{3.12} = 1.44$

Querying Phase

Steps

i. Create candidate pairs
ii. Symbol sequence
iii. Compute similarity of $\langle v, q\rangle$
iv. Greedy expansion

1. Compute similarity $\left(\chi^2_{\langle \text{v, q}\rangle}\right)$

$$\chi^2_{\langle \text{v, q}\rangle} = \sum_{\forall i} \frac{\left[O_{\langle \text{v, q}\rangle}(i) - E_{\langle \text{v, q}\rangle}(i)\right]^2}{O_{\langle \text{v, q}\rangle}(i)}$$

2. $O_{\langle \text{v, q}\rangle}(i)$ - Observed count of σ_i; $E_{\langle \text{v, q}\rangle}(i)$ - Expected count of σ_i
3. $O_{\langle \text{v}_1, \text{q}_1\rangle} = \{\sigma_1, \sigma_2, \sigma_3, \sigma_4\}$; Length of $O_{\langle \text{v}_1, \text{q}_1\rangle} = 3 + 1 = 4$

- $E_{\langle \text{v, q}\rangle}(i) = l \cdot \Pr(\sigma_i)$
  - l = length of $O_{\langle \text{v, q}\rangle}$
  - $E_{\langle \text{v}_1, \text{q}_1\rangle}(1) = 4 \cdot 0.78 = 3.12$
  - $O_{\langle \text{v}_1, \text{q}_1\rangle}(1) = 1$
  - $\chi^2_{\langle \text{v}_1, \text{q}_1\rangle}(1) = \frac{(1-3.12)^2}{3.12} = 1.44$

$$\Pr(\sigma_i) = \frac{1}{2}\left[\frac{1}{(1+(i-1)\cdot\kappa)^2} - \frac{1}{(1+i\cdot\kappa)^2}\right], \qquad 2 \le i \le \tau$$

$$\Pr(\sigma_1) = 1 - \sum_{j=2}^{\tau} \Pr(\sigma_j)$$

$$\Pr(\sigma_1) = 1 - 0.22 = 0.78$$

FIGURE 4: PROBABILITY CALCULATION

Querying Phase

Steps

i. Create candidate pairs
ii. Symbol sequence
iii. Compute similarity of $\langle v, q\rangle$
iv. Greedy expansion

1. Compute similarity $\left(\chi^2_{\langle \text{v, q}\rangle}\right)$

$$\chi^2_{\langle \text{v, q}\rangle} = \sum_{\forall i} \frac{\left[O_{\langle \text{v, q}\rangle}(i) - E_{\langle \text{v, q}\rangle}(i)\right]^2}{O_{\langle \text{v, q}\rangle}(i)}$$

2. $O_{\langle \text{v, q}\rangle}(i)$ - Observed count of σ_i; $E_{\langle \text{v, q}\rangle}(i)$ - Expected count of σ_i
3. $O_{\langle \text{v}_1, \text{q}_1\rangle} = \{\sigma_1, \sigma_2, \sigma_3, \sigma_4\}$; Length of $O_{\langle \text{v}_1, \text{q}_1\rangle} = 3 + 1 = 4$

- $E_{\langle \text{v, q}\rangle}(i) = l \cdot \Pr(\sigma_i)$
  - l = length of $O_{\langle \text{v, q}\rangle}$

- $E_{\langle \text{v}_1, \text{q}_1\rangle}(1) = 4 \cdot 0.78 = 3.12$
- $O_{\langle \text{v}_1, \text{q}_1\rangle}(1) = 1$
- $\chi^2_{\langle \text{v}_1, \text{q}_1\rangle}(1) = \frac{(1-3.12)^2}{3.12} = 1.44$

$$\Pr(\sigma_i) = \frac{1}{2}\left[\frac{1}{(1 + (i-1)\cdot\kappa)^2} - \frac{1}{(1 + i\cdot\kappa)^2}\right], \qquad 2 \le i \le \tau$$

$$\Pr(\sigma_1) = 1 - \sum_{j=2}^{\tau} \Pr(\sigma_j)$$

$$\Pr(\sigma_1) = 1 - 0.22 = 0.78$$

FIGURE 4: PROBABILITY CALCULATION

Querying Phase

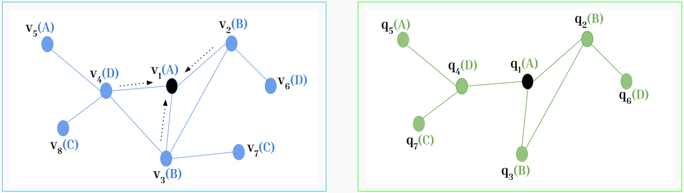
Steps

i. Create candidate pairs

ii. Symbol sequence

iii. Compute similarity of $\langle v, q \rangle$

iv. Greedy expansion



(A) (B)

FIGURE 4: EXAMPLE (A) TRAGET GRAPH ($\mathcal{G}$) AND (B) QUERY GRAPH ($\mathcal{Q}$)

- Expand the candidate vertex pair greedily based on χ^2 similarity.

– Neighbour pairs of $\langle v_1, q_1 \rangle$: $\{\langle v_2, q_2 \rangle, \langle v_3, q_3 \rangle, \langle v_4, q_4 \rangle\}$

– Choose vertex pair with highest similarity - $\langle v_4, q_4 \rangle$

– Add it to the candidate answer - $\{\langle v_1, q_1 \rangle, \langle v_4, q_4 \rangle\}$

– Repeat (explore neighbors of candidate answer) - $\{\langle v_2, q_2 \rangle, \langle v_3, q_3 \rangle, \langle v_5, q_5 \rangle, \langle v_8, q_7 \rangle\}$

Experimental Setup

Dataset	# Vertices	# Edges	# Unique Labels
Human	4,674	86,282	44
HPRD	9,460	37,081	307
Protein	43,471	81,044	3
Flickr	80,513	5.9M	195
IMDb	428,440	1.7M	22

TABLE 1: REAL-WORLD DATASETS

- Human, HPRD, Protein: Biological Networks
- Flickr: Social Interaction
- IMDb: Knowledge Graph

Query and Accuracy:

- 6 Types of Queries: exact + 5 noisy
- Noisy: edge addition/ deletion, vertex addition/deletion, modified label
- Query vertex sizes: 3, 5, 7, 9, 11, 13 (20 each)
- Total queries: $6 \times 6 \times 20 = 720$
- Accuracy: fraction of edges of $\mathcal{Q}$ present in answer

Performance

Dataset / Algorithm	Accuracy				
	Human	HPRD	Protein	Flickr	IMDb
VELSET [1]	0.42	0.65	0.37	0.75	0.53
G-Finder [2]	0.45	0.12	0.47	out of memory	
VERSACHI	**0.90**	**0.81**	**0.67**	**0.84**	**0.87**

TABLE 2: OVERALL ACCURACY

- Accuracy averaged across query types and sizes.
- $> 20\%$ accuracy improvements
- VERSACHI: substantial accuracy gain against slight increase in compute time

[1] Dutta et al. 2017, WWW. 1281–1290.

[2] Liu et al. 2019. IEEE BigData. 513–522.

Performance

Performance on Barabási-Albert graphs ($|V|$ = 50K, Avg. Degree = 50, κ = 0.001)

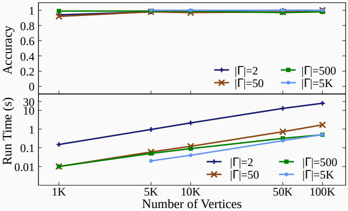


FIGURE 5: EFFECT OF GRAPH SIZE

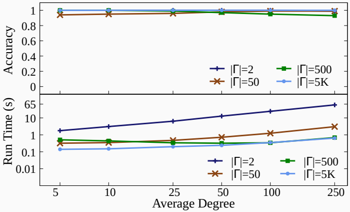


FIGURE 6: EFFECT OF AVERAGE DEGREE

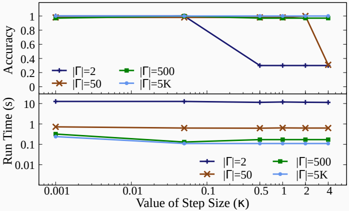


FIGURE 7: EFFECT OF STEP-SIZE (κ)

To summarize . . .

- Chebyshev's inequality - underlying graph distribution
- Statistical significance - capture deviations
- High accuracy across datasets and noisy queries
- VERSACHI : approximate labelled graph querying
  – scalable and accurate

The source code of VERSACHI is available from https://github.com/shubhangiat/VerSaChI

Thank you !!
Questions?