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

Shubhangi Agarwal[a,1], Sourav Dutta[b,2], Arnab Bhattacharya[a,3]

[a]Indian Institute of Technology Kanpur, India , [b]Huawei Research Centre, Dublin, Ireland

Problem Statement

Given, a simple, deterministic, undirected, vertex labeled graphs
$G = (V_G, E_G, L_G)$ and
$Q = (V_Q, E_Q, L_Q)$
Find top-k subgraphs of G that are *best* matching (maximum similarity) to Q in terms of vertex labels and edge overlap.

Contributions

VERSACHI: an approximate subgraph querying method which

- Is highly **scalable**
- Captures **2-hop similarity**
- Models graph chracteristics using **Chebyshev's inequality**
- Returns **statistically significant** answers

VERSACHI

Offline Phase - for input graph G

1. Create Invertex Label index (IL_G) and Neighbor Label index (NL_G).
   - $IL_G(A) : \{v_1, v_5\}$; – $NL_G(v_1) : \{A, B, B, D\}$
2. Calculate neighborhood overlap for all vertex pairs, $\langle u, v\rangle \in G \times G$.
   - Modified Tversky index, with penalizing factor $\gamma = 3$

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

   - $\eta_{\langle v_1, v_4\rangle} = 3/(3 + 2^3) = 3/11$
3. Compute characteristics of G.
   - Avg. neighbor overlap score; $\psi(G) = 0.59$
   - Std. dev. of overlap scores; $\delta(G) = 0.38$
   - Max. z-score of overlap score; $\Delta(G) = 1.54$
4. Discretize into category symbols.
   - Step size, κ, lower values preferred, e.g., $\kappa = 0.1$
   - #Category symbols, $\tau = \lceil(\Delta(G) - 1)/\kappa\rceil$; $\tau = \lceil(1.54 - 1)/0.1\rceil = 6$
   - Symbols $\sigma_{i \in [2,\tau]}$ - defined for κ-sized ranges.
   - σ_1 defined for range $[0, 1 + \kappa]$, negative values as well
   - $\eta_{v_1, v_4} = -0.83$; Category: σ_1

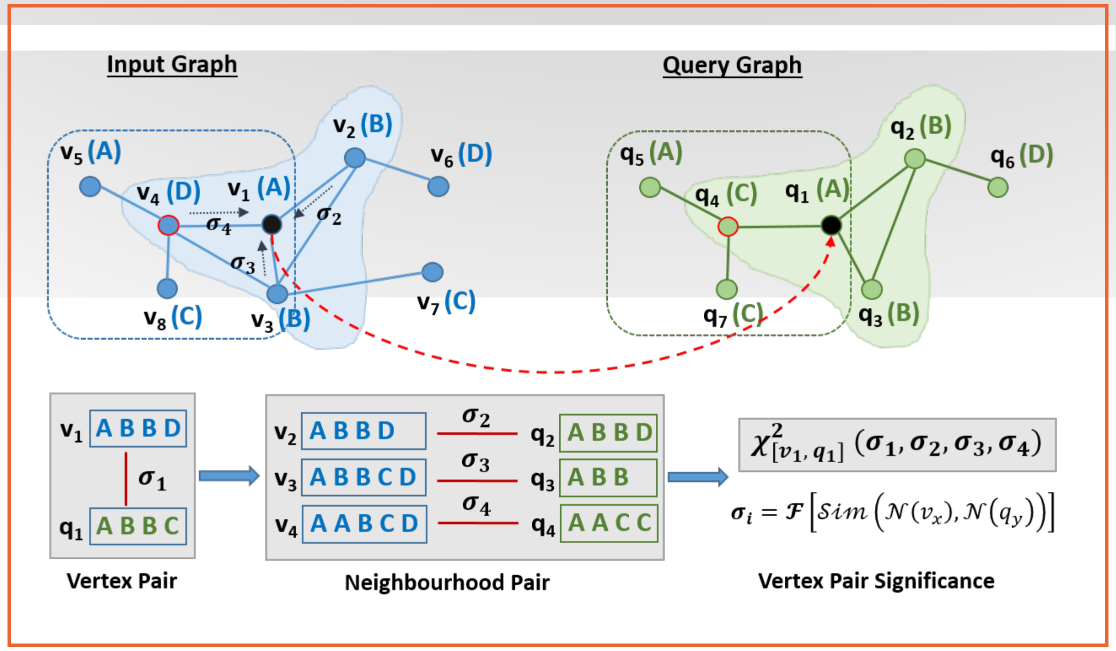

Figure: Example Graphs

5. Compute probability for each symbol (Chebyshev's inequality model).
   - $\Pr(\sigma_{i\in[2,\tau]}) = \frac{1}{2}\left[\frac{1}{(1+(i-1)\cdot\kappa)^2} - \frac{1}{(1+i\cdot\kappa)^2}\right]$; $\Pr(\sigma_1) = 1 - \sum_{j=2}^{\tau}\Pr(\sigma_j)$
   - $\Pr(\sigma_1) = 1 - 0.22 = 0.78$

Querying Phase - for each query Q

i. Generate indices (IL_Q, NL_Q) and candidate pairs based on label match.
   - $\langle v_1, q_1\rangle$, $\langle v_5, q_5\rangle$, $\langle v_8, q_7\rangle$ etc.

ii. Compute observed vertex symbol sequence.
   - Neighbor overlap score of vertex pair $\langle v, q\rangle$
   - Overlap scores of neighbor vertex pairs of $\langle v, q\rangle$ - greedy mapping
   - $O_{\langle v_1, q_1\rangle} = \{\sigma_1, \sigma_2, \sigma_3, \sigma_4\}$

iii. Compute statistical significance of candidate pairs.
   - Expected value, $E_{\langle v,q\rangle}(\sigma_i) = len(O_{\langle v,q\rangle}) \cdot \Pr(\sigma_i)$
   - $\chi^2_{\langle v,\, q\rangle} = \sum_{\forall i}\left[O_{\langle v,\, q\rangle}(\sigma_i) - E_{\langle v,\, q\rangle}(\sigma_i)\right]^2 / O_{\langle v,\, q\rangle}(\sigma_i)$
   - $O_{\langle v_1, q_1\rangle}(\sigma_1) = 1$; $E_{\langle v_1, q_1\rangle}(\sigma_1) = 4 \cdot 0.78 = 3.12$

iv. Expand greedily, prefer neighbor pairs with high χ^2 values
   - Maintain a priority queue of neighbor pairs of candidate answer subgraph
   - Expand till list exhausts or subgraph same as query
   - Priority queue for $\{\langle v_1, q_1\rangle, \langle v_4, q_4\rangle\} : \{\langle v_2, q_2\rangle, \langle v_3, q_3\rangle, \langle v_5, q_5\rangle, \langle v_8, q_7\rangle\}$

Experimental Results

- Evaluated on 5 real world datasets and 720 exact and noisy queries of 6 different sizes
- VERSACHI showed >20% accuracy improvement over baselines on average

Dataset / Algorithm	Human	HPRD	Protein	Flickr	IMDb
	Accuracy				
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: Overall Accuracy

- Linear increase in runtime with increase in graph size and density for Barabási-Albert graphs while accuracy remains largely unaffected

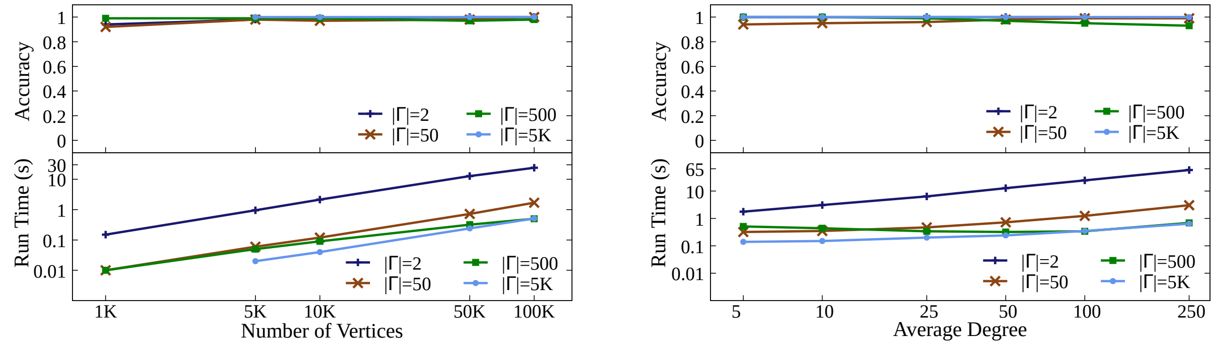

Figure: Effect of (a) Graph size and (b) Average degree ($|V| = 50K$, Avg. Degree = 50, $\kappa = 0.001$)

- Accuracy falters for larger step sizes (κ); Higher $\kappa \Rightarrow$ Number of symbols decrease
- Lower number of symbols limit the power of VERSACHI to differentiate between finer differences in neighborhood mismatches between graphs

References

[1] *VELSET: Dutta et al. 2017, WWW. 1281–1290;*
[2] G-Finder: *Liu et al. 2019. IEEE BigData. 513–522.*

For more details

Scan the QR codes.

⟵ **arXiv paper** | **github code** ⟶

arXiv

GitHub

If you have any questions, you can reach out to us!

[1]sagarwal@cse.iitk.ac.in, [2]sourav.dutta2@huawei.com
[3]arnabb@cse.iitk.ac.in