VENOM: Approximate Subgraph Matching with Enhanced Neighbourhood Structural Information

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

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

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

CODS COMAD 2024
IIIT Bangalore, India

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

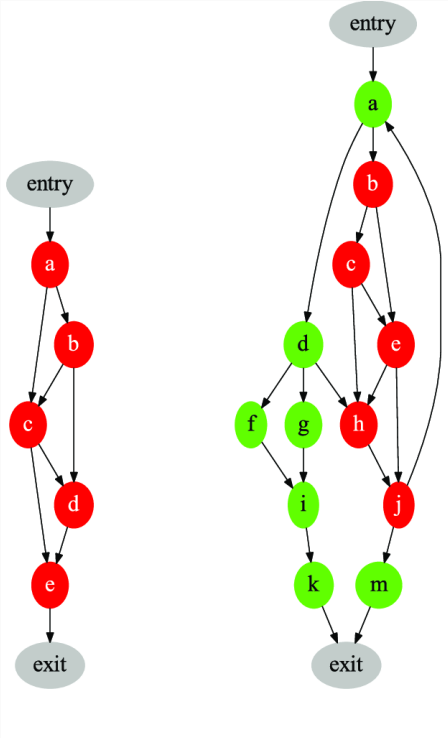
[††]Huawei Research Centre, Dublin, **Ireland**

Approximate Subgraph Matching (ASM)

Objective: Search in a graph dataset for subgraph ***similar*** to the pattern of interest.

Approximate Subgraph Matching (ASM)

Objective: Search in a graph dataset for subgraph ***similar*** to the pattern of interest.

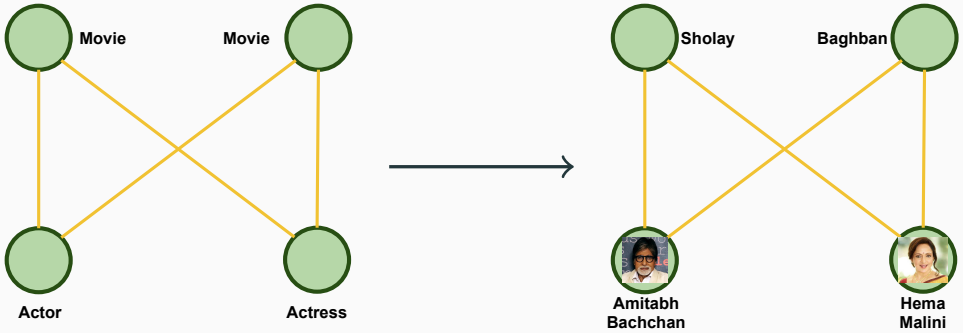


Malware Detection

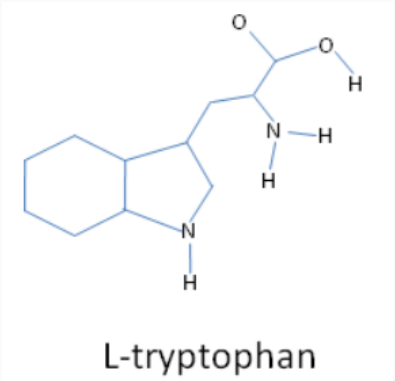
with Control Flow Graph representations

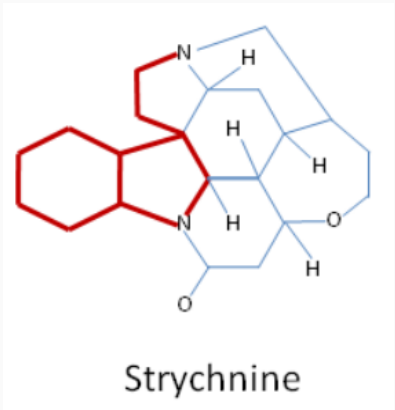
of a Program

Source:
https://ieeexplore.ieee.org/document/6838703



Question-Answering in Knowledge Graphs





Compound Analysis

Source: https://www.cs.tau.ac.il/~roded/sigma.pdf

Goal of Study

Factors affecting ASM performance:

- Number of nodes
- Density of graph
- Label distribution
- Query size

⋮

- Size of neighborhood matched
- Degree distribution

Goal of Study

Factors affecting ASM performance:

- Number of nodes
- Density of graph
- Label distribution
- Query size

⋮

- **Size of neighborhood matched**
- Degree distribution

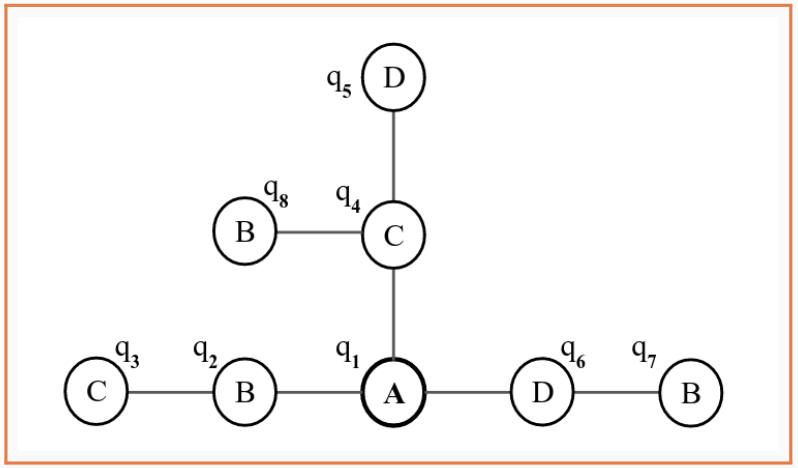
Goal of Study

Factors affecting ASM performance:

- Number of nodes
- Density of graph
- Label distribution
- Query size

⋮

- **Size of neighborhood matched**
- Degree distribution



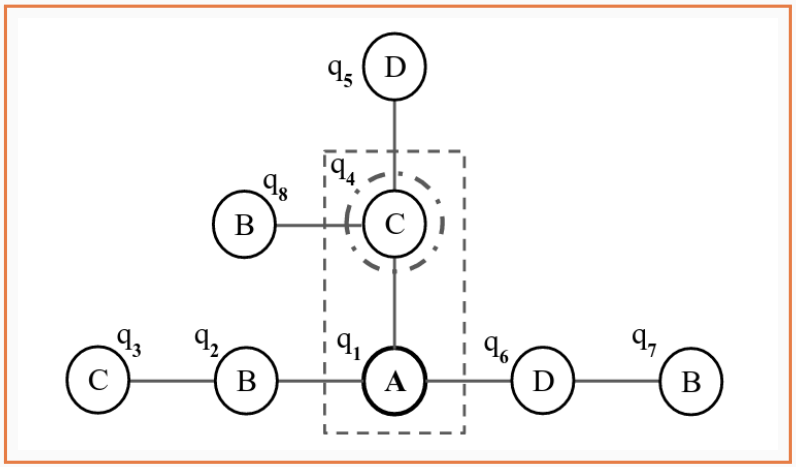
EXAMPLE GRAPH

→ **Depth of neighborhood (h)**

→ **Breadth of neighborhood (k)**

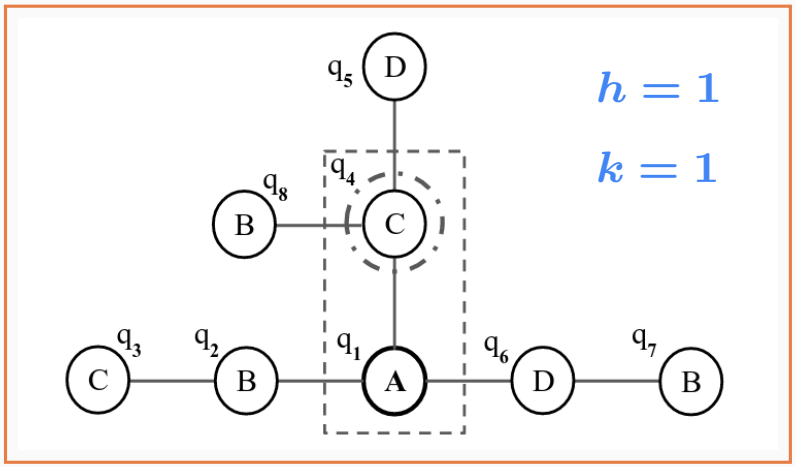
VENOM

Parametrizing size of neighborhood $h \implies$ **depth** $k \implies$ **breadth**



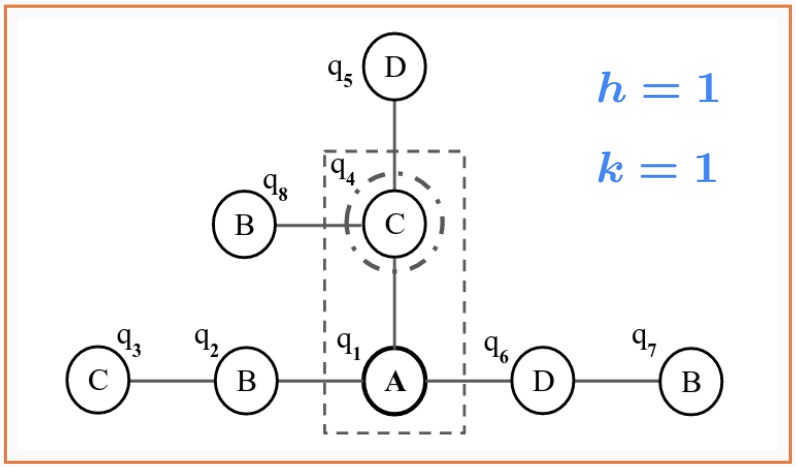
VENOM

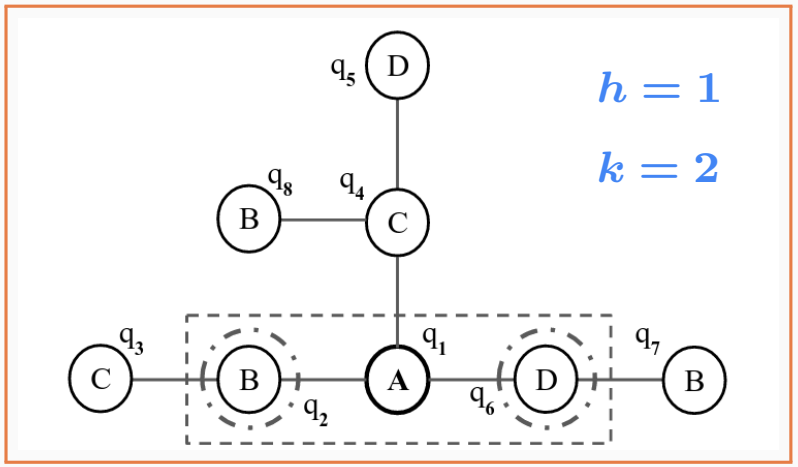
Parametrizing size of neighborhood $h \implies$ **depth** $k \implies$ **breadth**



VENOM

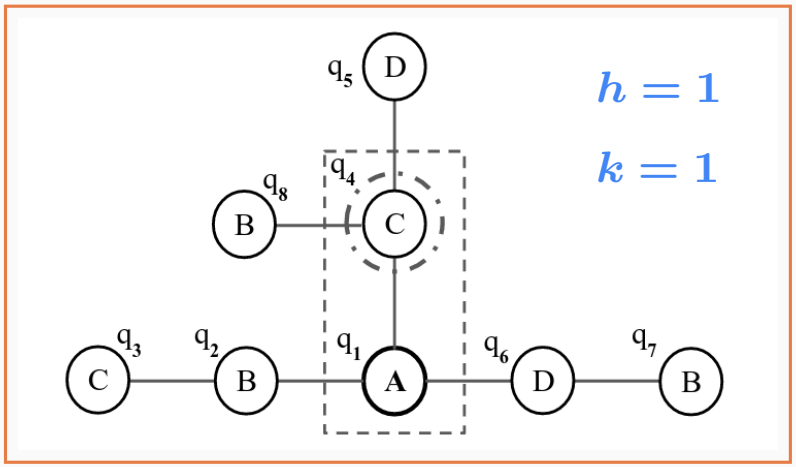
Parametrizing size of neighborhood $h \implies$ **depth** $k \implies$ **breadth**

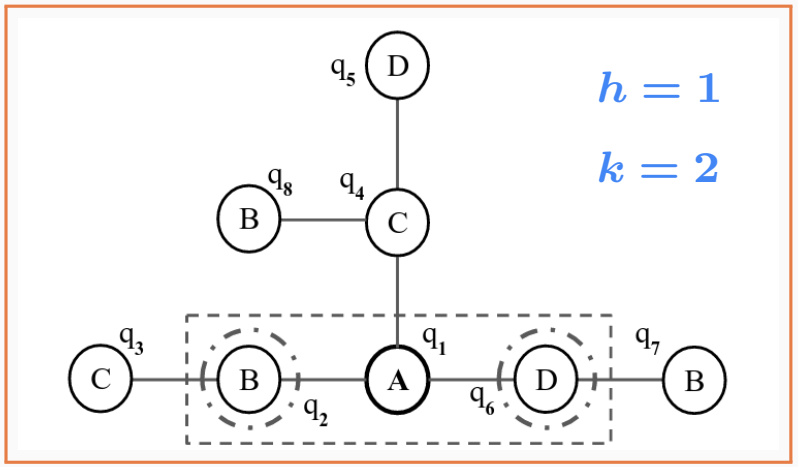


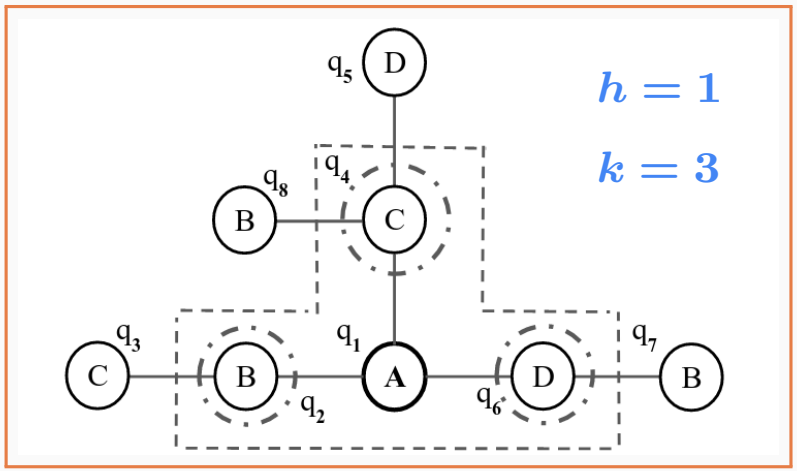


VENOM

Parametrizing size of neighborhood $h \implies$ **depth** $k \implies$ **breadth**

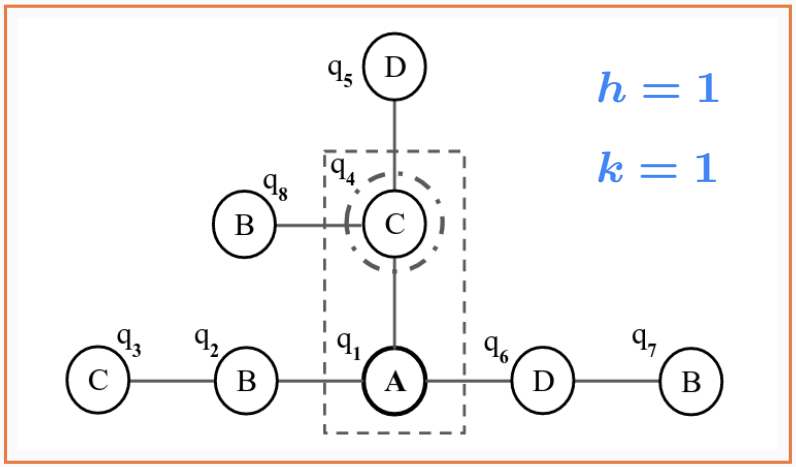


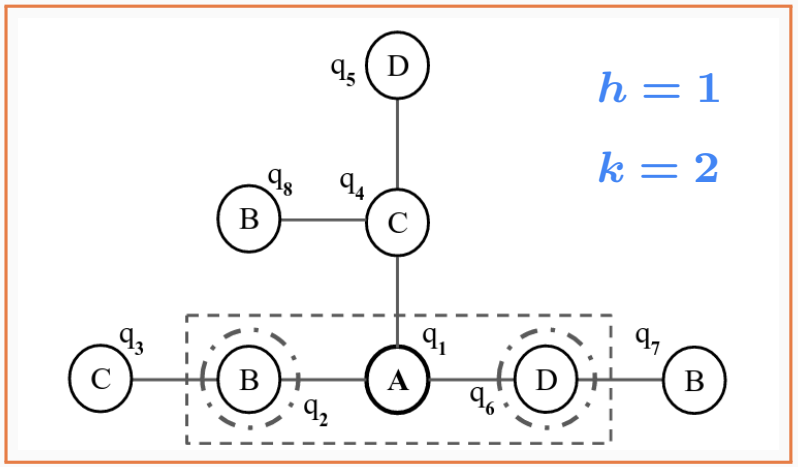


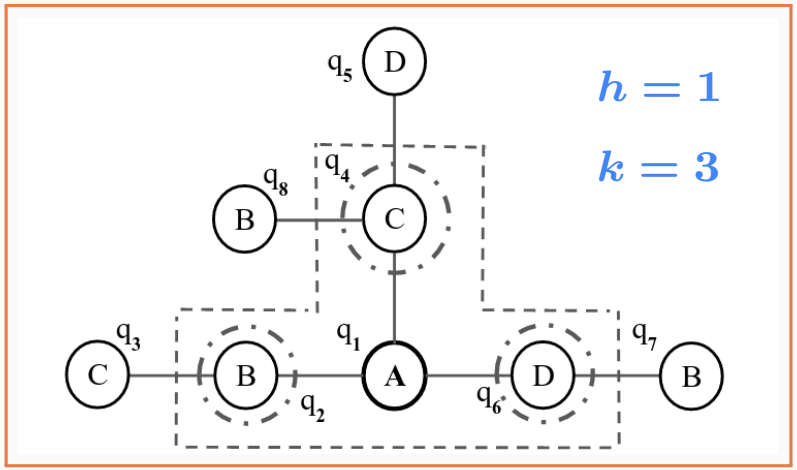


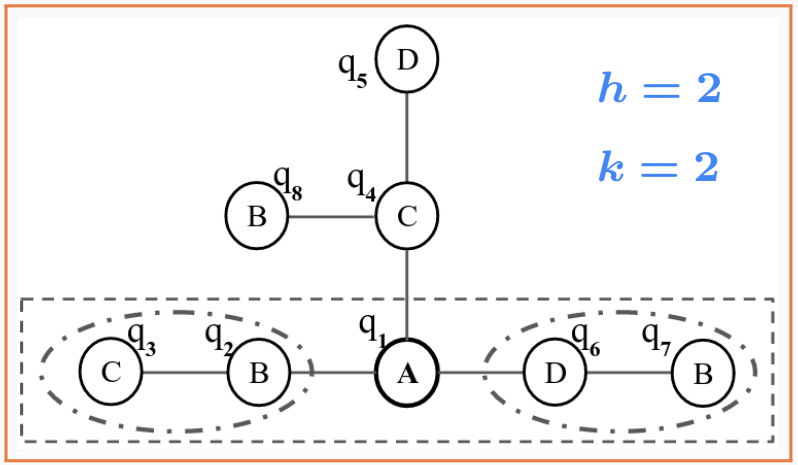
VENOM

Parametrizing size of neighborhood $h \implies$ **depth** $k \implies$ **breadth**









Units and Groups

VENOM-(k, h)

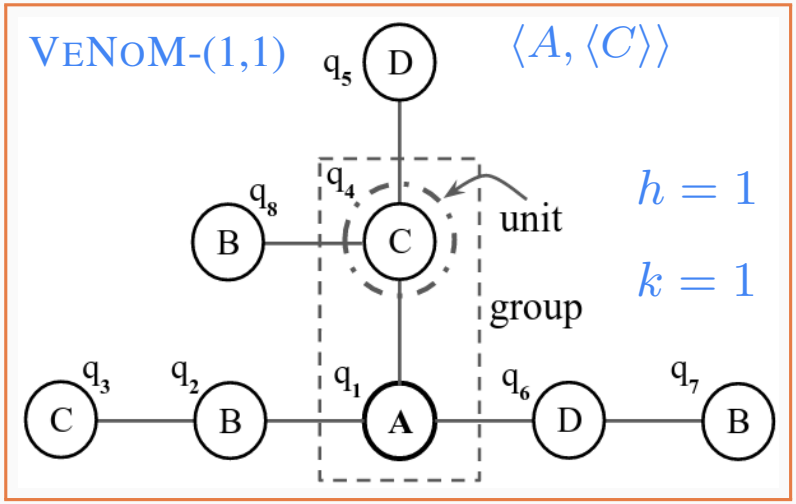
Different instances of VENOM based on breadth (k) and depth (h) of the neighborhood.

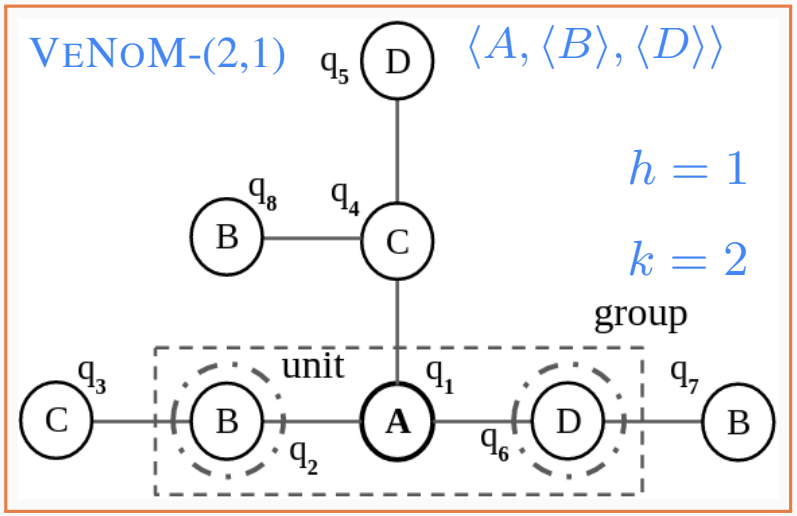
Unit

An ordered collection of neighbor labels of a vertex that forms a path of length h.

Group

A set of k units of a vertex along with its label where the units correspond to different neighbors.

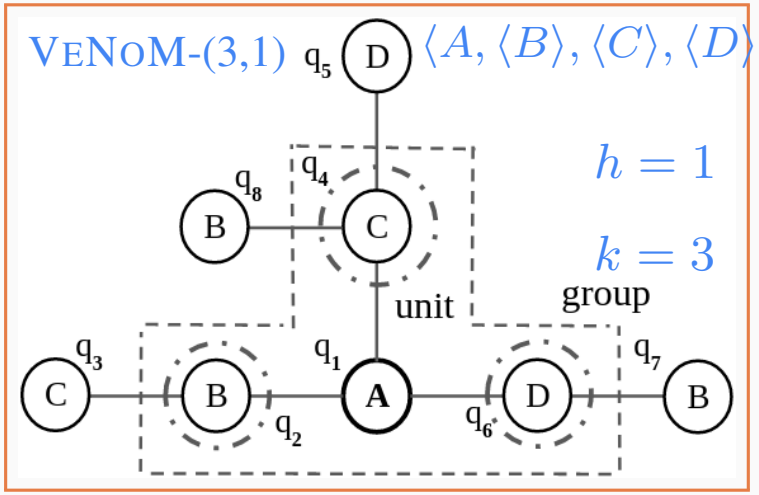




Units and Groups

VENOM-(k, h)

Different instances of VENOM based on breadth (k) and depth (h) of the neighborhood.



Unit

An ordered collection of neighbor labels of a vertex that forms a path of length h.

Group

A set of k units of a vertex along with its label where the units correspond to different neighbors.

Units and Groups

VENOM-(k, h)

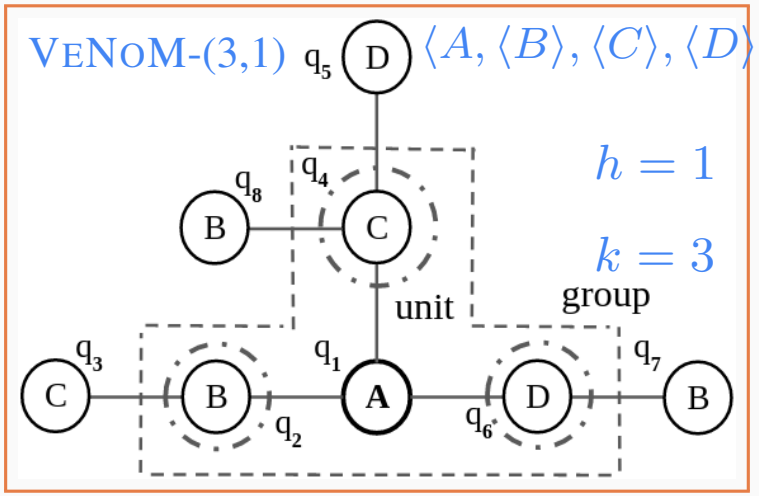
Different instances of VENOM based on breadth (k) and depth (h) of the neighborhood.

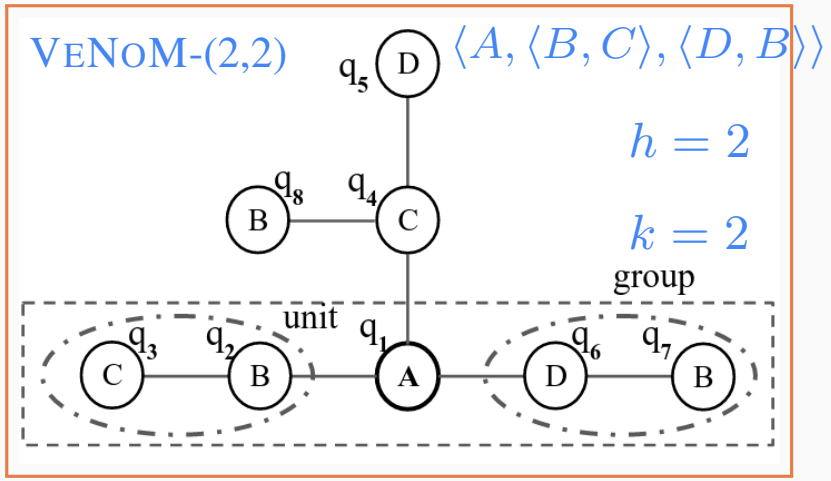
Unit

An ordered collection of neighbor labels of a vertex that forms a path of length h.

Group

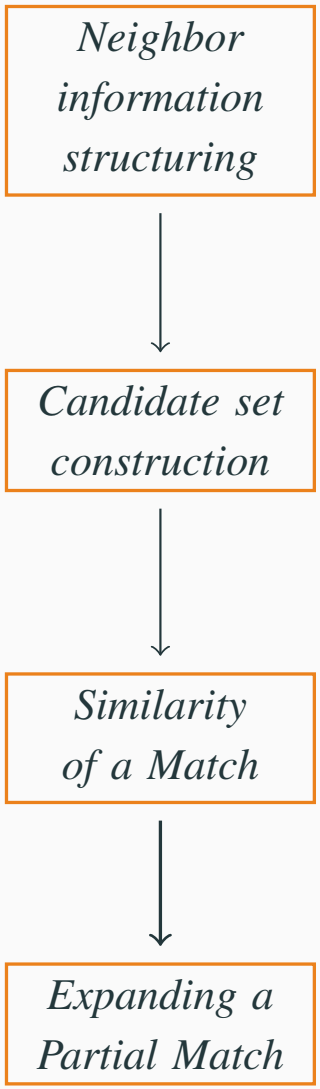
A set of k units of a vertex along with its label where the units correspond to different neighbors.





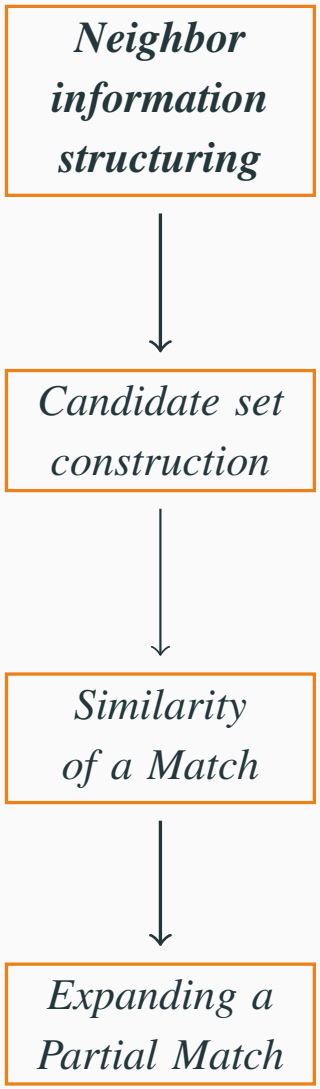
VENOM

VENOM STAGES



VENOM

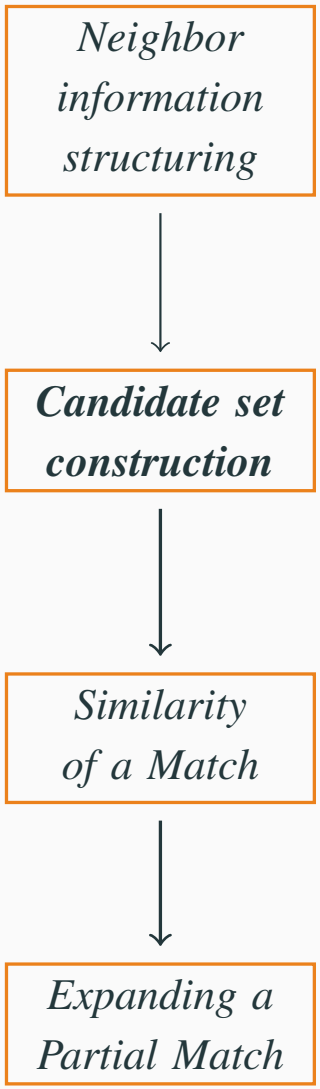
VENOM STAGES



1. Number of unique labels in target graph $\mathcal{G}$
2. Degree of each vertex $v \in \mathcal{G}$
3. Enumerate *units* and *groups* of v

VENOM

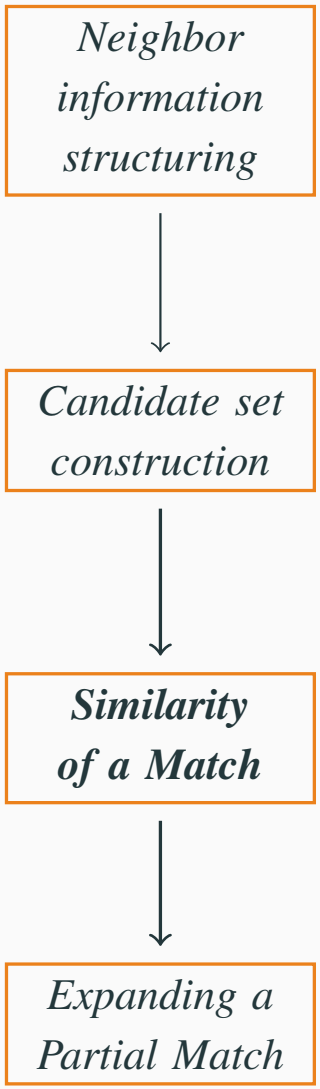
VENOM STAGES



- Candidate vertex: target vertices with label same as query node label
- Label similarity: e.g., Jaccard similarity

VENOM

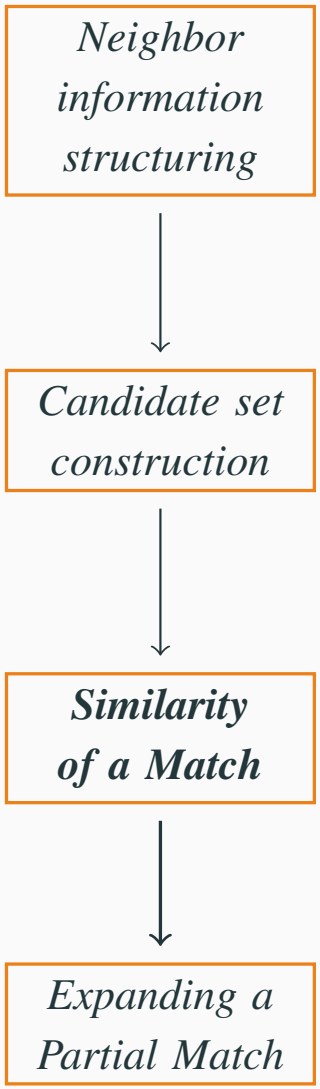
VENOM STAGES



- Goal: find most statistically significant match
- VENOM-(2,2): 5 group match categories
  - Ex: $\langle \underline{\text{A}}, \langle \text{B}, \text{C} \rangle, \langle \text{D}, \text{B} \rangle \rangle$
  - s_0: $\langle \underline{\text{A}}, \langle \times, \times \rangle, \langle \times, \times \rangle \rangle$
  - s_1: $\langle \underline{\text{A}}, \langle \text{B}, \times \rangle, \langle \times, \times \rangle \rangle$
  - s_2: $\langle \underline{\text{A}}, \langle \text{B}, \times \rangle, \langle \times, \text{B} \rangle \rangle$, $\langle \underline{\text{A}}, \langle \text{B}, C \rangle, \langle \times, \times \rangle \rangle$
  - s_3: $\langle \underline{\text{A}}, \langle \text{B}, \times \rangle, \langle \text{D}, \text{B} \rangle \rangle$
  - s_4: $\langle \underline{\text{A}}, \langle \text{B}, \text{C} \rangle, \langle \text{D}, \text{B} \rangle \rangle$
- $s_i \succ s_j, \qquad$ if $i > j$
- Enumerate groups and compare
  - Ordered comparison based on label overlap
  - Categorize group matches (s_i)
  - Large overlap preferred

VENOM

VENOM STAGES



- Similarity metric
  - Deviation of observed behavior from expected
  - Pearson χ^2 statistic

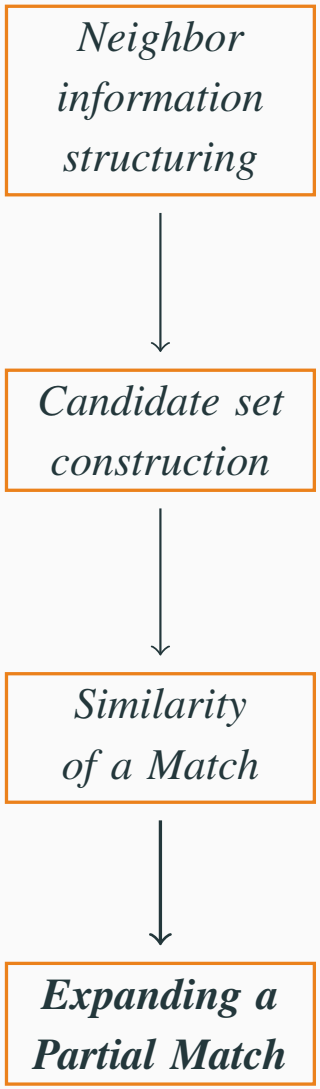
$$\chi^2 = \sum_i \frac{[O(s_i) - E(s_i)]^2}{E(s_i)}$$

  - Random vertex pair $\rightarrow$ Low similarity
  - Higher deviation $\implies$ Exceptional similarity

- Expected behavior: based on
  - Number of unique labels
  - Degree of the node
  - Number of groups of query node

VENOM

VENOM STAGES



1. Greedy expansion
2. Seed: candidate pair with highest χ^2
3. Prefer unmatched neighbors with highest χ^2
4. Repeat
   – Until terminal condition met

Setup

Datasets

REAL-WORLD DATASETS

Dataset	#Vertices	#Edges	#Labels
Human	4.6K	86.2K	44
HPRD	9.4K	37K	307
Flickr	80.5K	5.9M	195
PPI	12K	10.74M	2.4K

SYNTHETIC BARABÁSI-ALBERT GRAPHS
(200 DIFFERENT GRAPHS)

(#Vertices,
- $2K$
- $10K$
- $50K$
- $250K$
- $1000K$

Density,
- 4
- 10
- 40
- 100
- 250

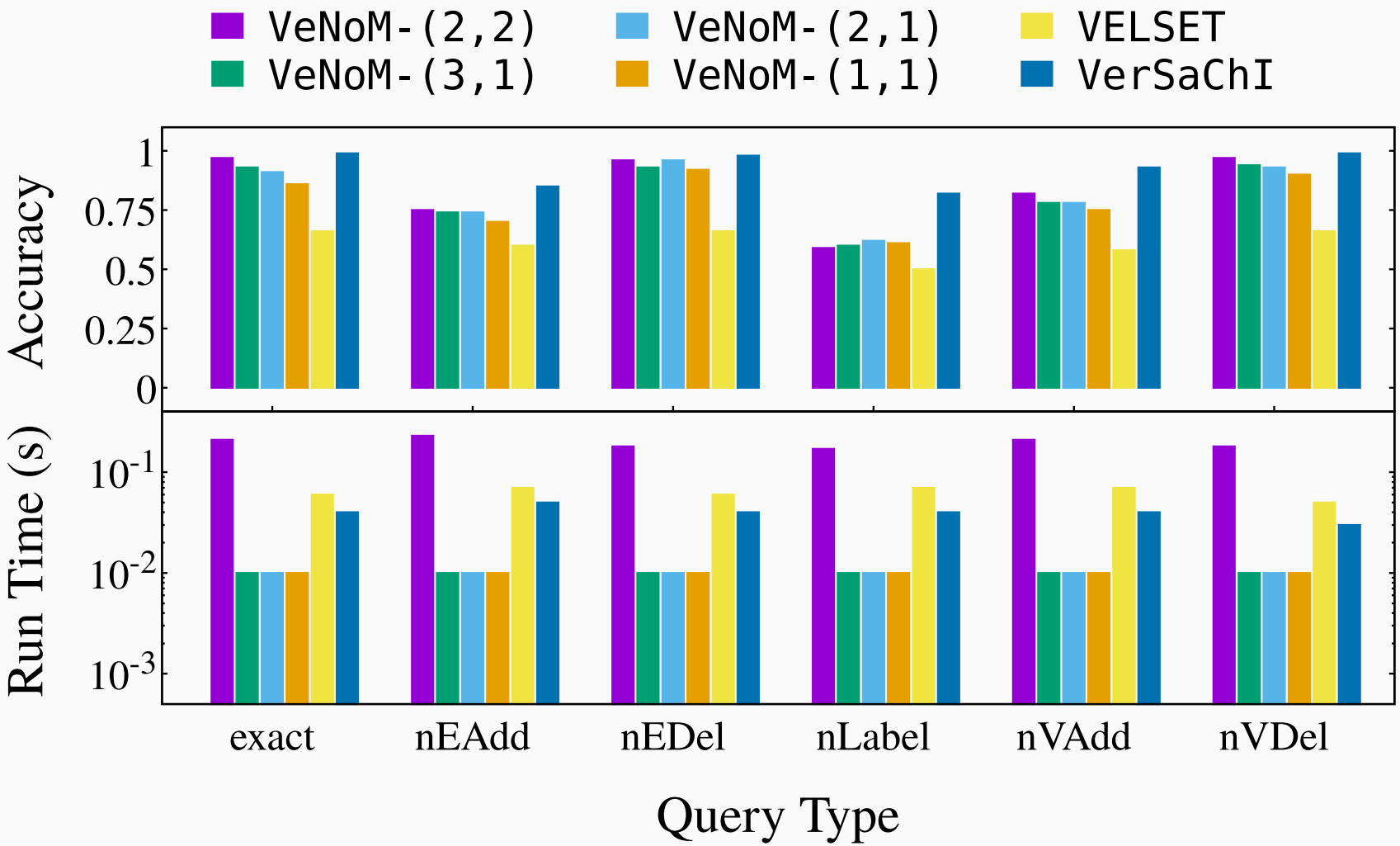
#Labels)
- 2
- 5
- 10
- 25
- 50
- 150
- 500
- $5K$

Query set

– Random connected subgraph extraction

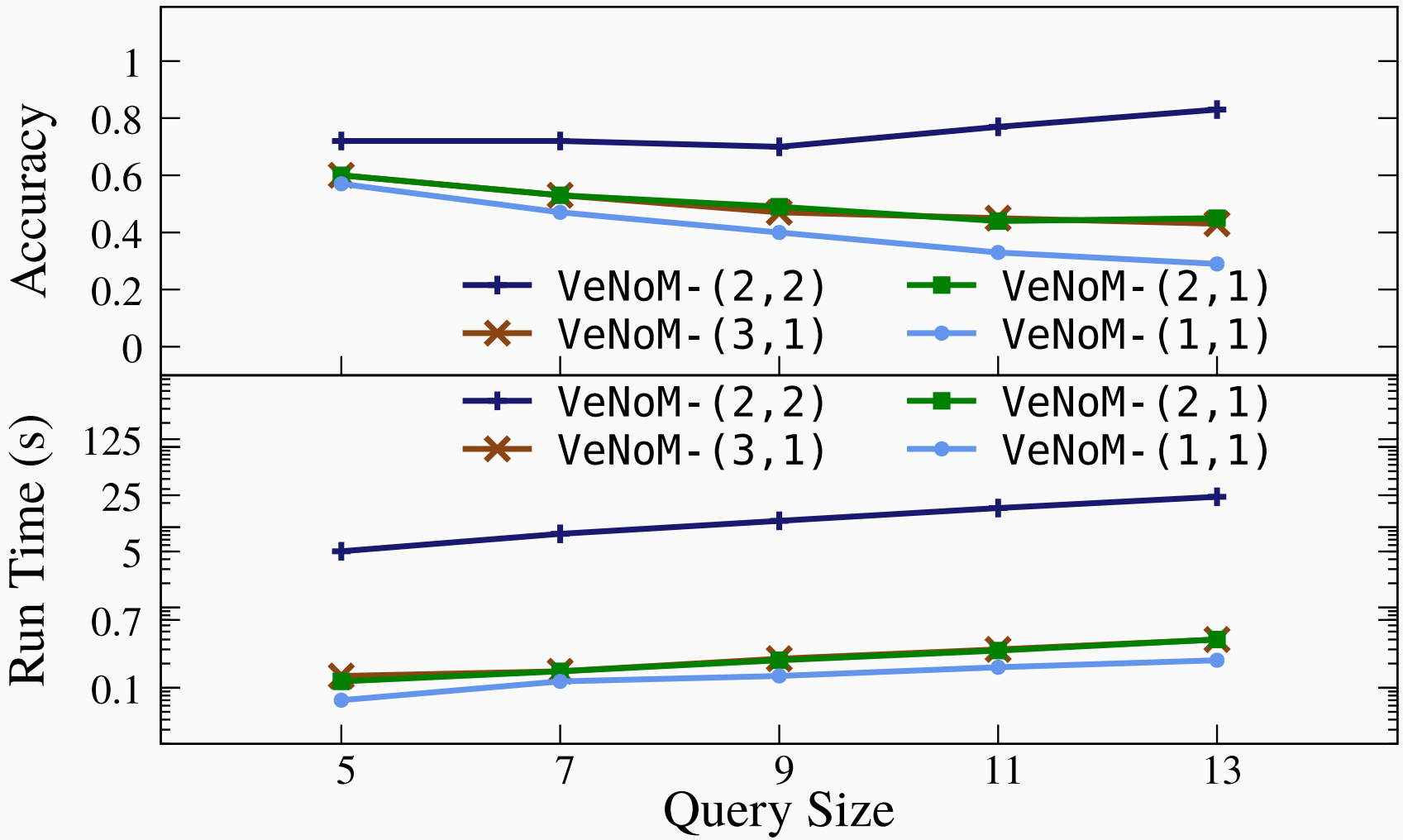
– Perturbations for noisy set

Real Graphs - HPRD



VELSET - Dutta *et al*, WWW 2017
VERSACHI- Agarwal *et al*, CIKM 2021

BA Graphs - Query Size



Results

- Increase in graph size and density
  - Accuracy decreases with increase in runtime
- Increase in label set
  - — Accuracy increases and runtime decreases
- VENOM-(2,2) overall more robust than counterparts
  - Higher accuracy due to structural look-ahead ability
  - Accuracy drop slower
- Limited improvement in accuracy
  - VENOM-(1,1), VENOM-(2,1), VENOM-(3,1)
- Runtime decreases with larger group size
  - For graphs with lower aveage degree

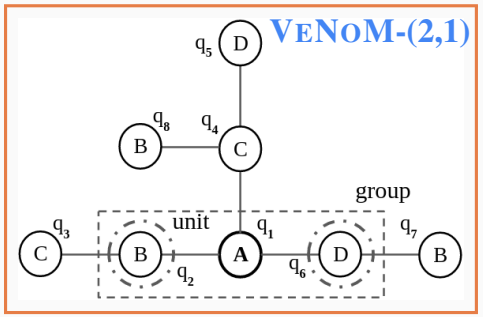
Summary

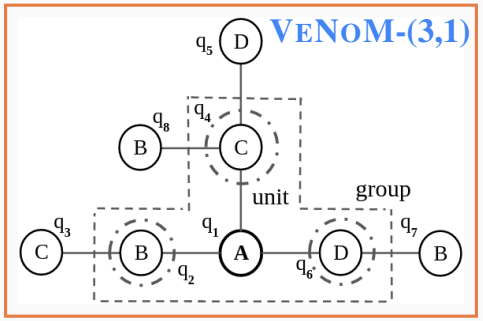
1. ***Key contribution***: *Units* and *Groups*

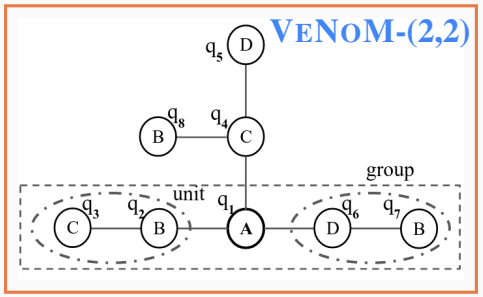
2. Effects of neighborhood size
   – Increase in depth $\Longrightarrow$ trade-off between time and accuracy

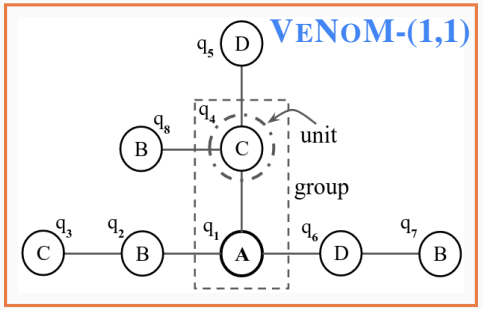
3. Runtime decreases with larger group size
   – Increase in breadth with depth may increase runtime efficiency

4. Diversified experiments over various parameters on both real and synthetic datasets
   – Size of graph, vertex degree, number of labels, query degree etc.









Questions?

Thank you!

Paper

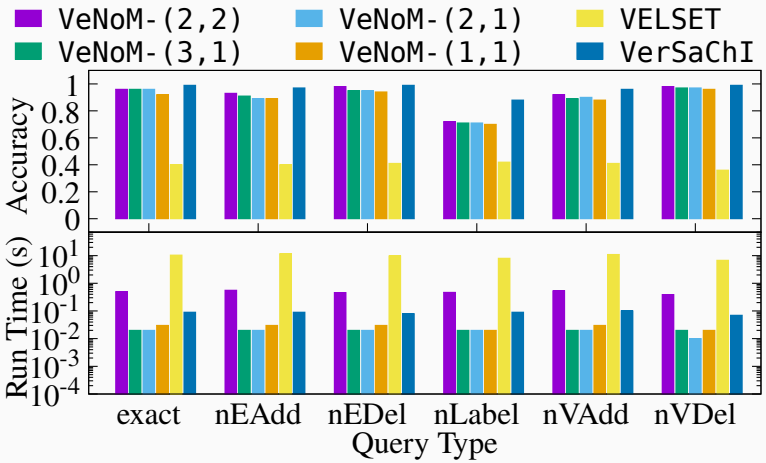
Code

Homepage

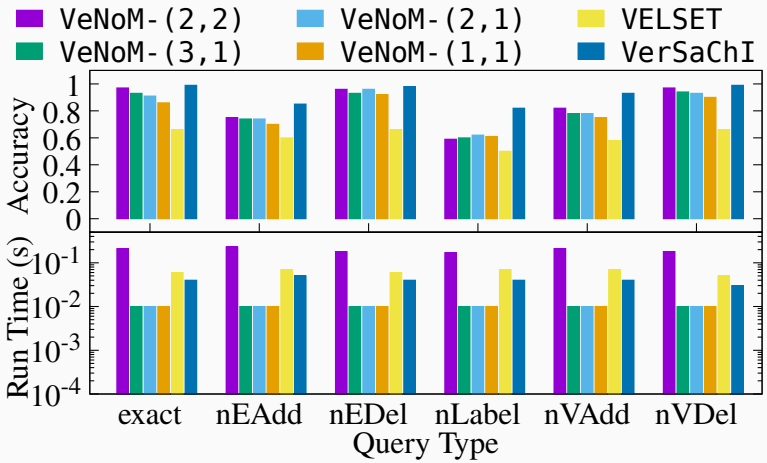
LinkedIn

Results: Real-world Graphs

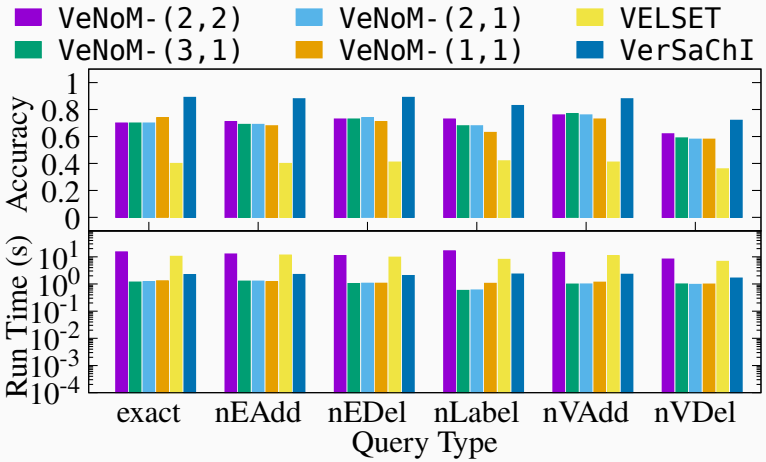
Human



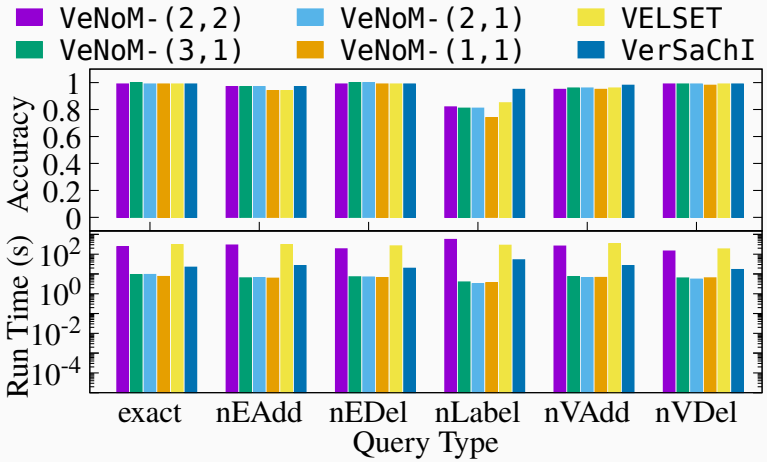
HPRD



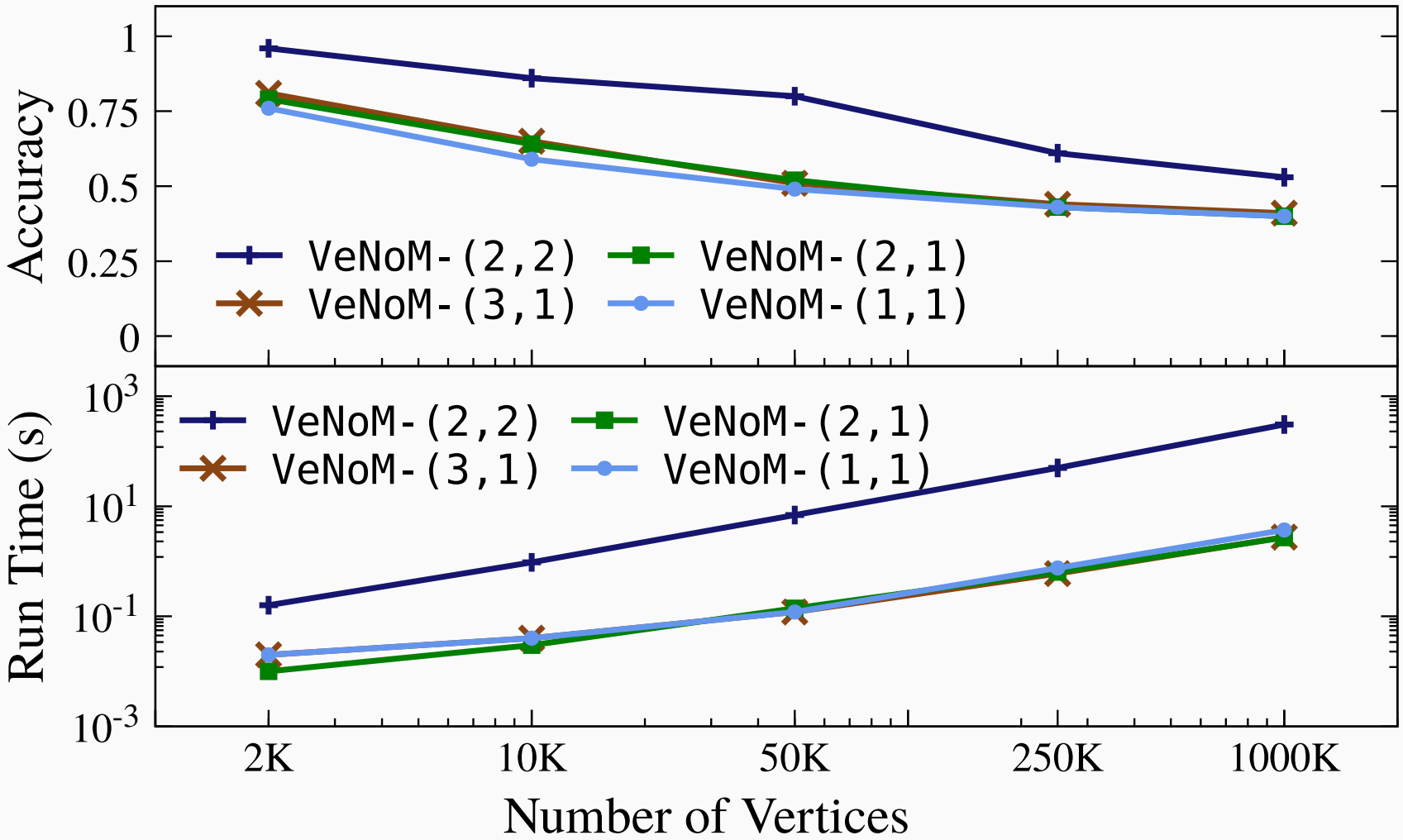
Flickr



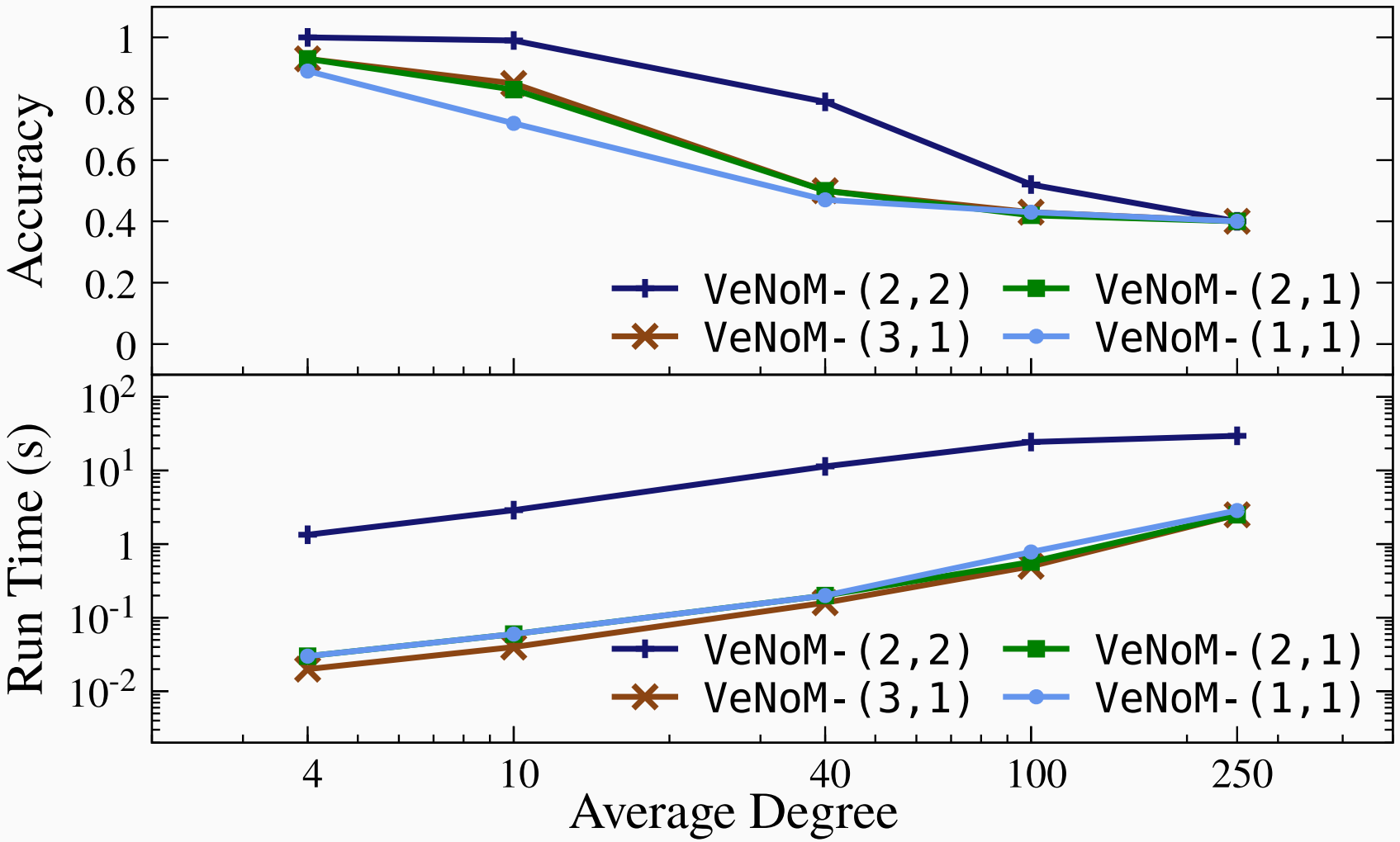
PPI



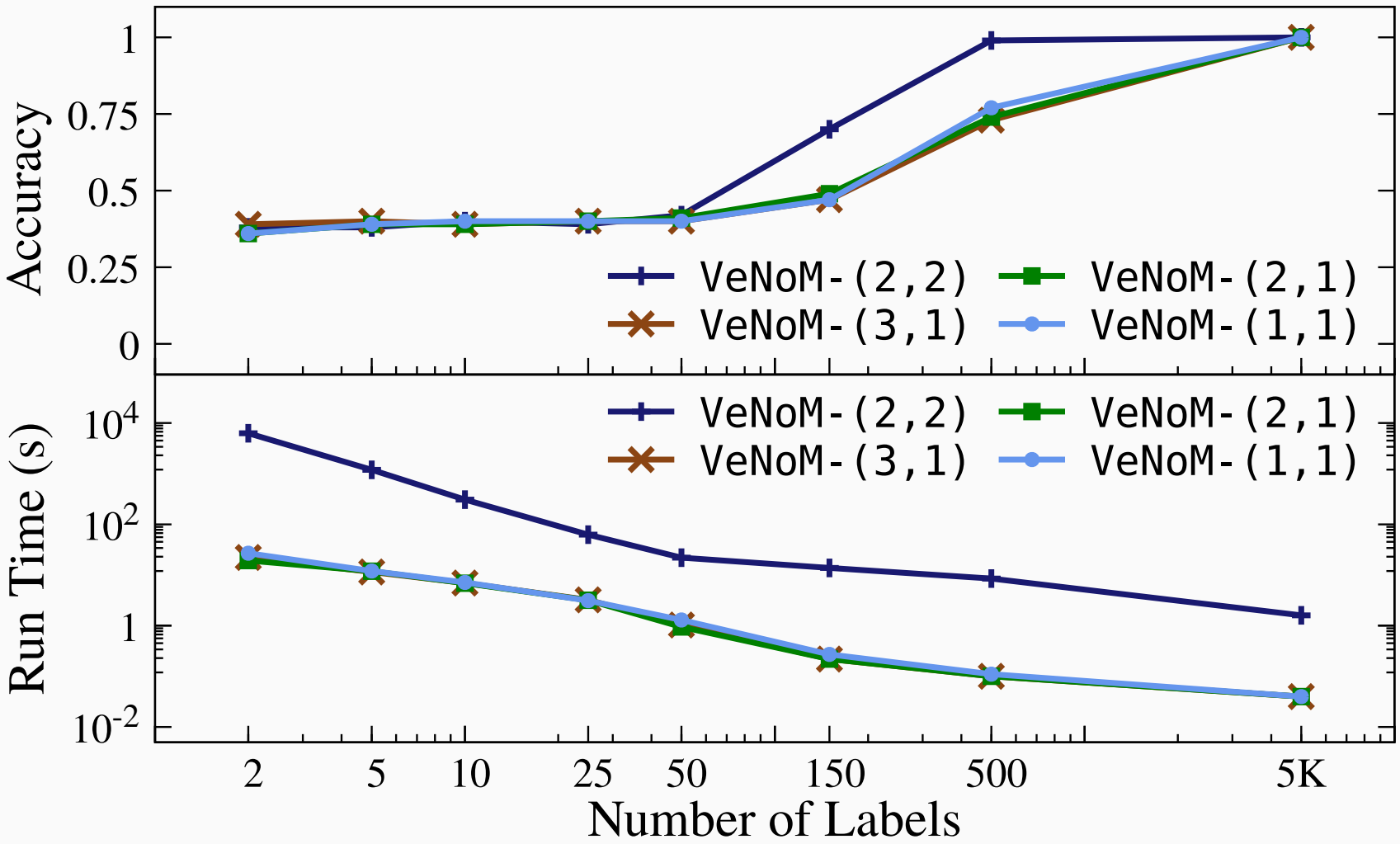
BA Graphs - Vertex Scaling



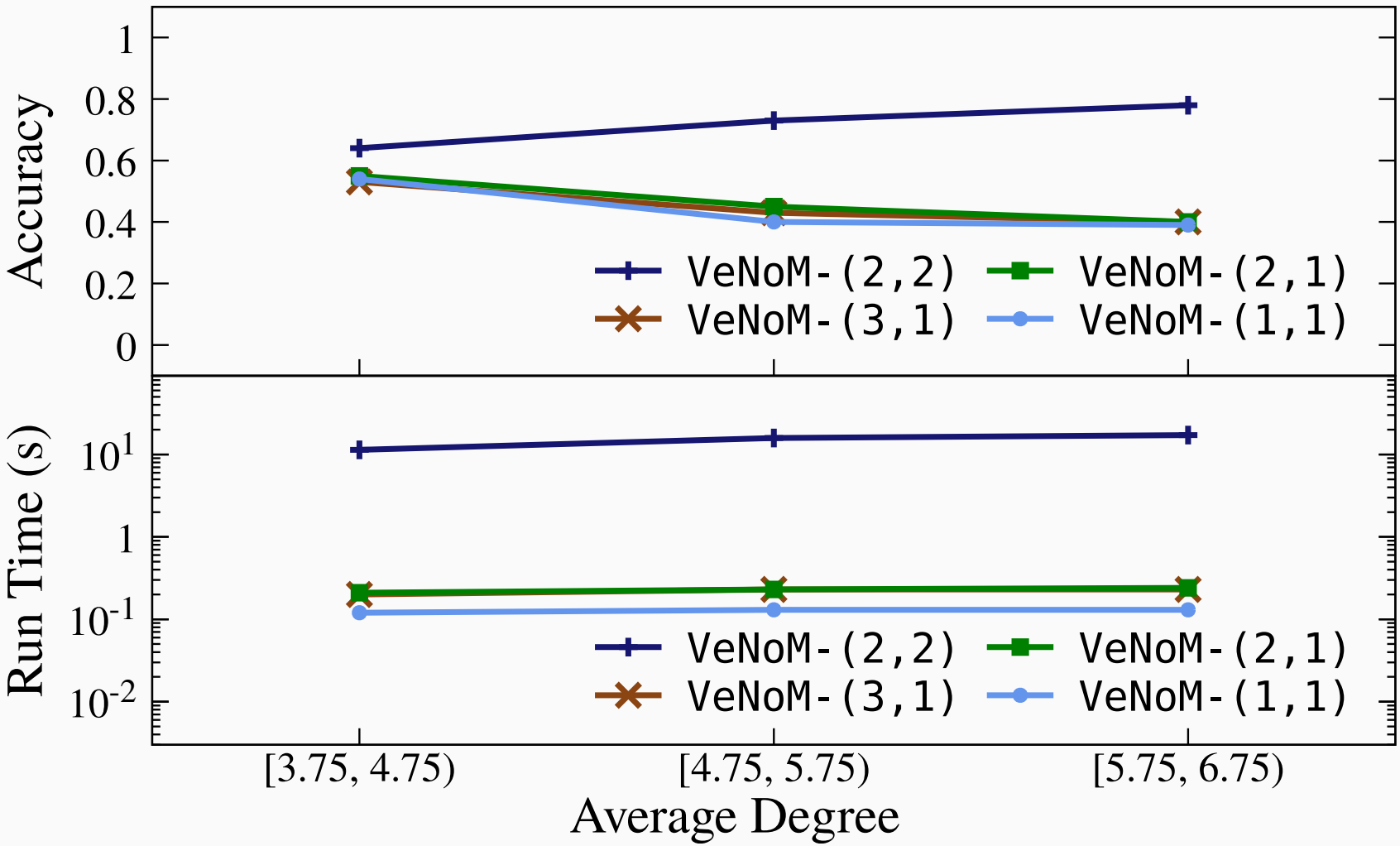
BA Graphs - Degree Scaling



BA Graphs - Label Scaling



BA Graphs - Query Degree



BA Graphs - Noisy Query