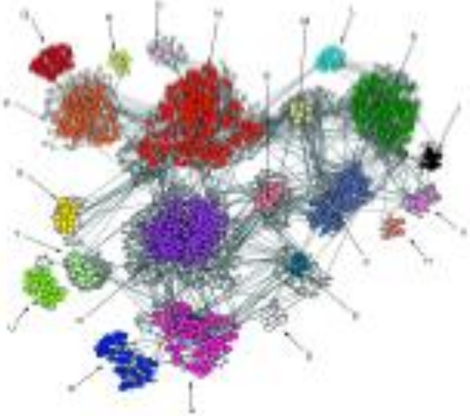
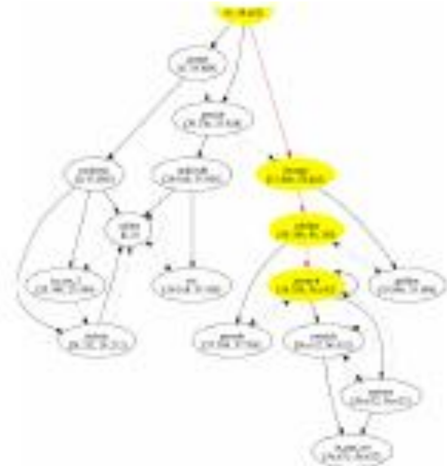
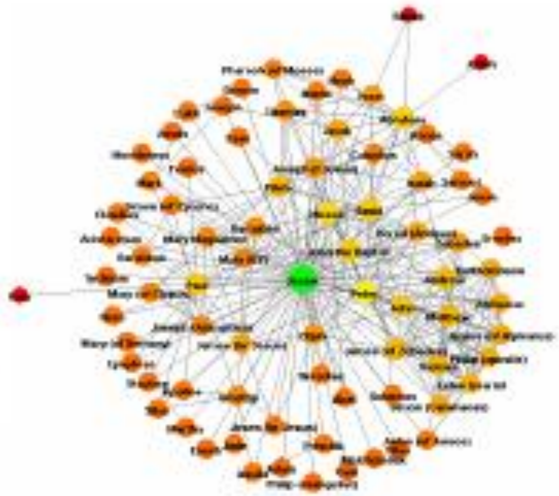
Mining Graph Patterns



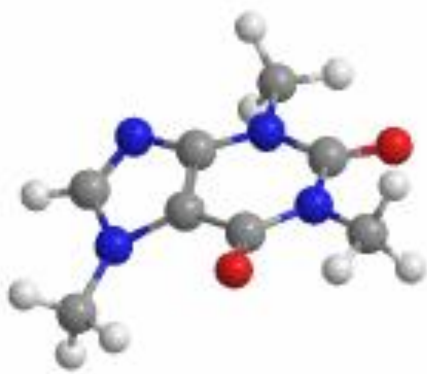
Co-expression Network



Program Flow



Social Network



Chemical Compound

Protein Structure

Why mine graph patterns?

- Direct Use:
  - Mining over-represented sub-structures in chemical databases
  - Mining conserved sub-networks
  - Program control flow analysis
- Indirect Uses:
  - Building block of further analysis
    - Classification
    - Clustering
    - Similarity searches
    - Indexing

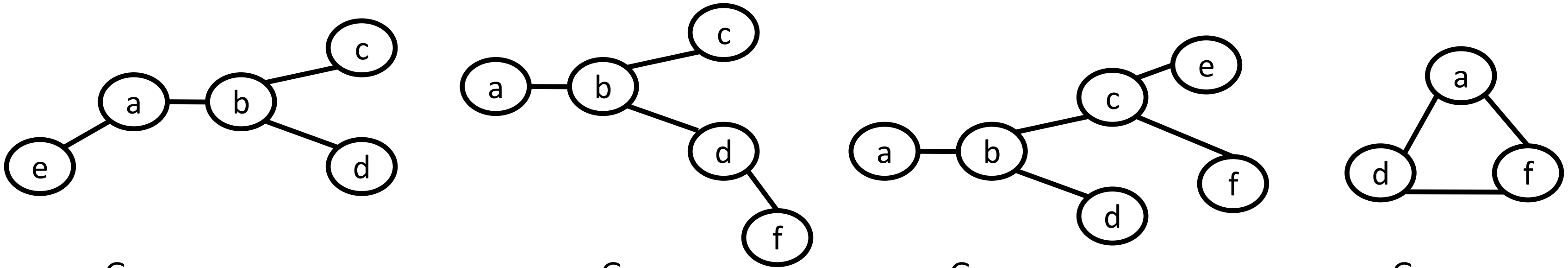
What are graph patterns?

- Given a function *f(g)* and a threshold θ, find all subgraphs *g*, such that *f(g)* ≥ θ.
- Example: frequent subgraph mining.

Given a graph dataset D, find subgraph g, s.t.

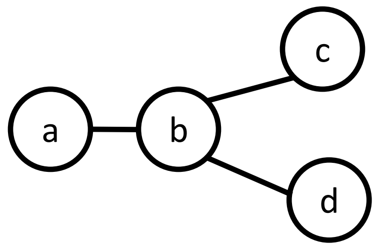
$$freq(g) \geq \theta$$

where $freq(g)$ is the percentage of graphs in D that contain g.



G_1 G_2 G_3 G_4

Θ=3



Frequent subgraph

Is this the only frequent subgraph?

NO!

Apriori Property

If a graph is frequent, all of its subgraphs are frequent.

Other Mining Functions

- Maximal frequent subgraph mining
  - A subgraph is *maximal*, if none of it super-graphs are frequent
- Closed frequent subgraph mining
  - A frequent subgraph is *closed*, if all its supergraphs have a lesser frequency
- Significant subgraph mining
  - G-test, p-value

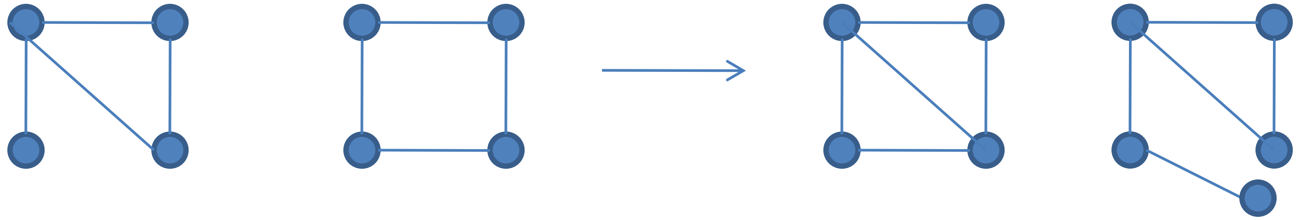
Frequent Subgraph Mining

- Apriori-based approach
  - AGM/AcGM: Inokuchi, et al. (PKDD'00)
  - FSG: Kuramochi and Karypis (ICDM'01)
  - PATH#: Vanetik and Gudes (ICDM'02, ICDM'04)
  - FFSM: Huan, et al. (ICDM'03) and SPIN: Huan et al. (KDD'04)
  - FTOSM: Horvath et al. (KDD'06)
- Pattern growth approach
  - Subdue: Holder et al. (KDD'94)
  - MoFa: Borgelt and Berthold (ICDM'02)
  - gSpan: Yan and Han (ICDM'02)
  - Gaston: Nijssen and Kok (KDD'04)

Frequent subgraph mining

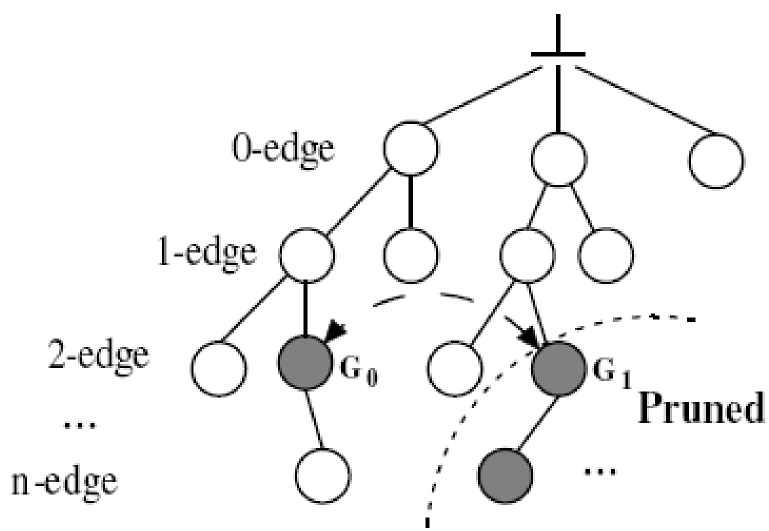
- Apriori Based Approach (FSG)
  - Find all frequent subgraphs of size K
  - Find candidates of size *k+1* edges by joining candidates of size *k edges*
  - Must share a common subgraph of *k-2* edges

  Example: (FSG)



Pattern Growth Approach

- Pattern Growth Approach
  - Depth first exploration
  - Recursively grow a frequent subgraph

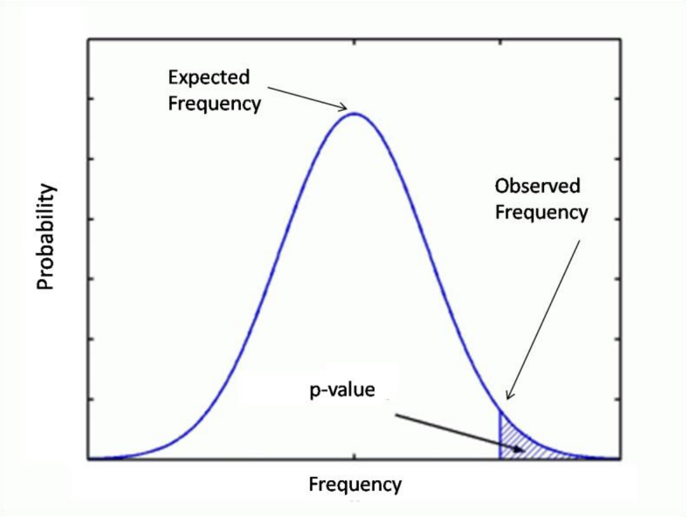


Mining Significant subgraphs

- What is significance?
  - Gtest, p-value
  - Both attempt to measure the deviation of the **observed frequency** from the **expected frequency**
  - Example: Snow in Santa Barbara is significant, but snow in Alaska is not.

P-value

- p-value : what's the probability of getting a result as extreme or more in the possible range of test statistics as the one we actually got?



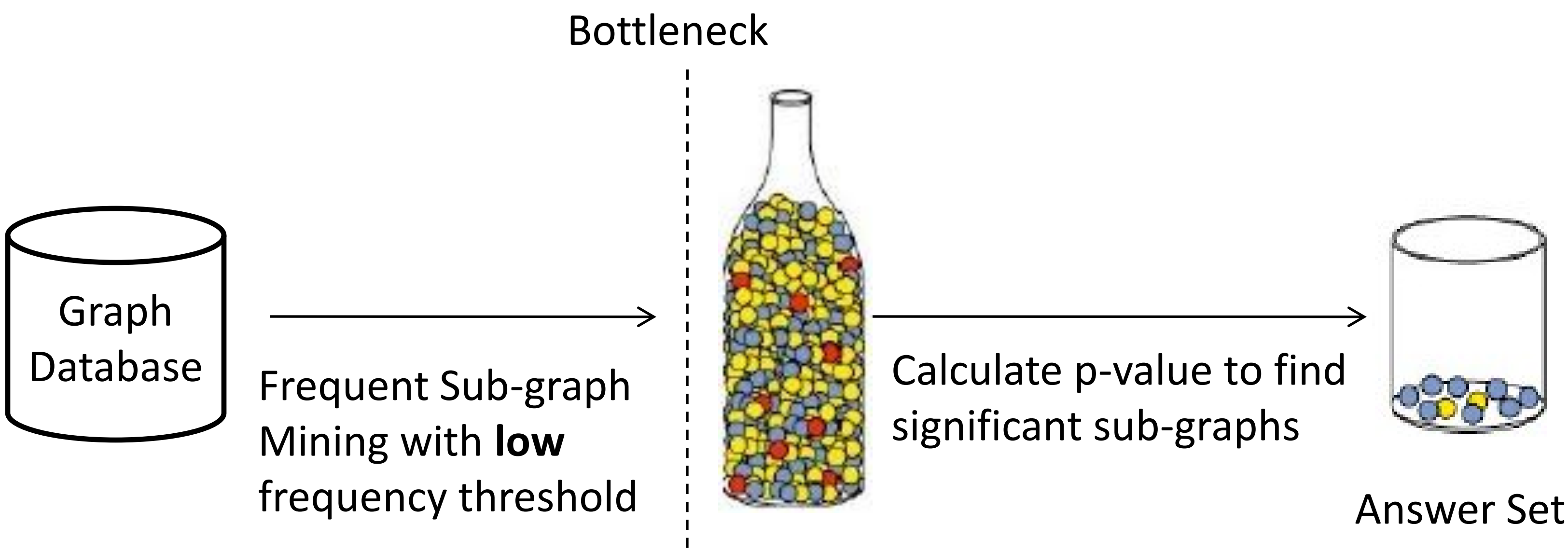
- Lower the p-value, higher the significance

Problem formulation

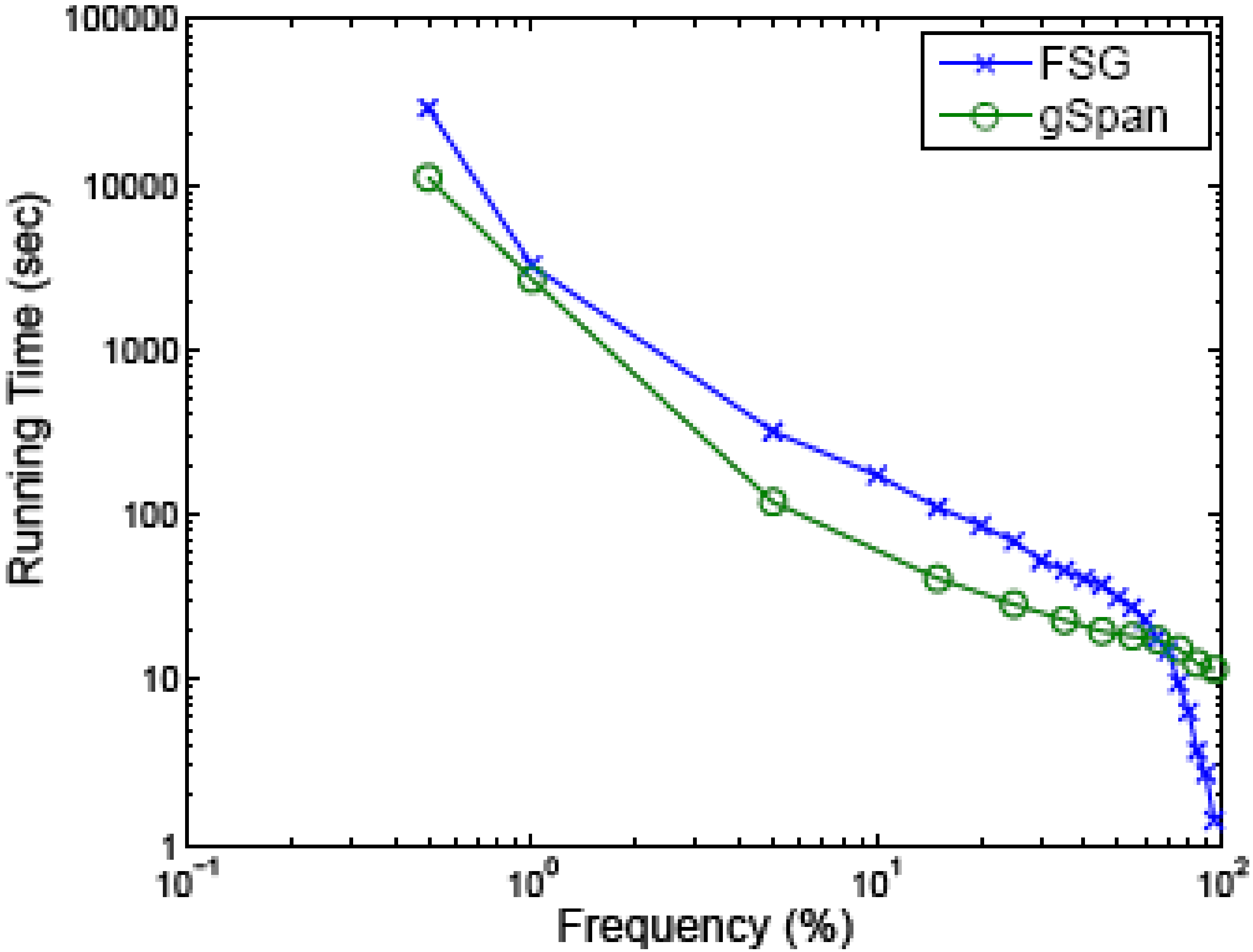
- Find answer set $\mathbb{A} = \{g | p\text{-}value(g) \leq \eta, g \subseteq G, G \in \mathbb{D}\}$
  - $\mathbb{D}$: Graph Database
  - **η** **:** Significance Threshold
  - $g \subseteq G$: *g* is a subgraph of *G*

- Low frequency does not imply low significance and vice versa
  - Graph with frequency 1% can be significant if expected frequency is 0.1%

Solution to Problem: Approach 1

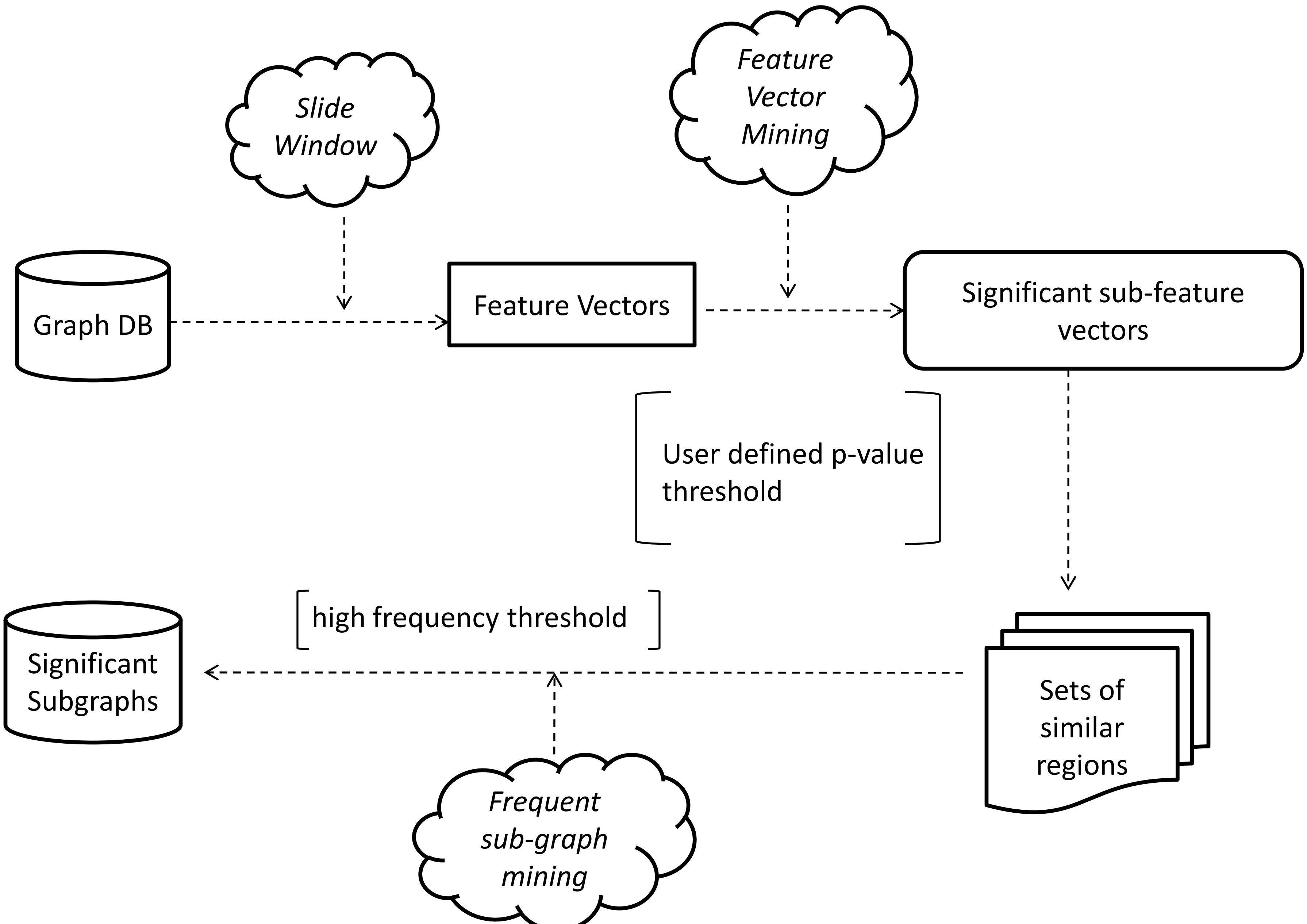


- Number of frequent subgraphs grow exponentially with frequency



FSG
gSpan
100000
10000
1000
100
10
1
Running Time (sec)
10^{-1}
10^{0}
10^{1}
10^{2}
Frequency (%)

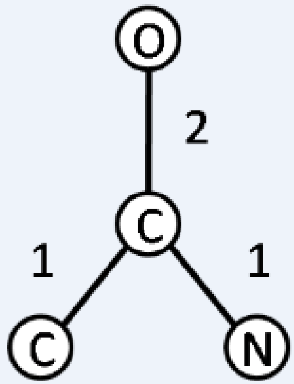
Alternative Approximate Solution



Converting graphs to feature vectors

- Random walk with Restart (RWR) on each node in a graph
- Feature vectors discretized to 10 bins

Graphical Representation



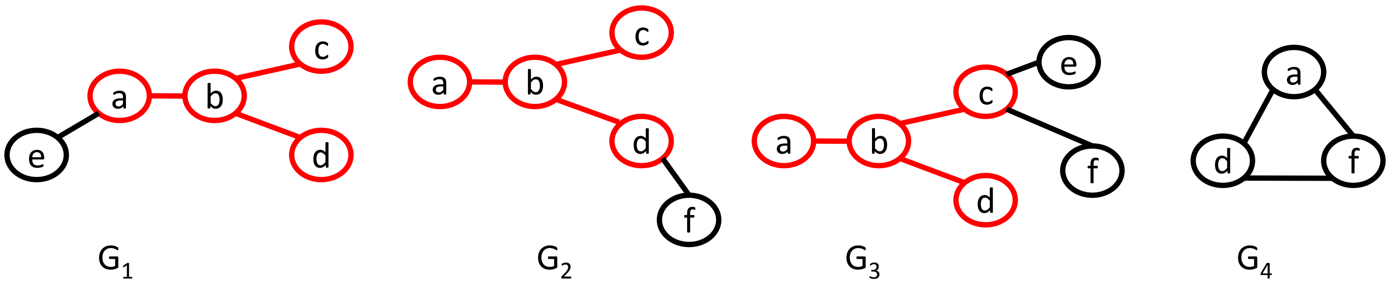
Random Walk Results

ID	Starting Atom	O-2-C	C-1-C	C-1-N
h_1	O	4	2	2
h2	C	2	3	3
h3	C	2	4	2
h4	N	2	2	4

What does RWR vectors preserve?

- Distribution of node-types around each node in graph
- Stores more structural information than a simple count of node-types
- Captures the feature vector representation of the subgraph around each node in a graph

Extracting information from feature vectors



Vector	a-b	a-d	a-e	a-f	b-c	b-d	c-e	c-f	d-f
G_1	2	0	3	0	1	1	0	0	0
G_2	4	0	0	0	2	1	0	0	1
G_3	3	0	0	0	1	2	1	1	0
G_4	0	3	0	3	0	0	0	0	2

- Floor of G_1,G_2,G_3: [2,0,0,0,1,1,0,0,0]
- Floor of G_1,G_2,G_3,G_4 : [0,0,0,0,0,0,0,0,0]
- False positives pruned later

Measuring p-value of feature vector

- Sub-feature vector: $\underline{X}=[x_1, ..,x_n]$ is a sub-feature vector of $\underline{Y}=[y_1, ..,y_n]$ if $x_i \le y_i$ for $i=1..n$.
  - Example: [2,3 1] ≤ [4,3,2].
  - In other words, "$\underline{X}$ occurs in $\underline{Y}$"
- Given a vector $\underline{X}$:
  - $P(X)$ = *Probability of* $\underline{X}$ occurring in an arbitrary $\underline{Y}$

$$= P(y_1 \ge x_1, ..,y_n \ge x_n)$$

$$= \prod_{i=1}^{n} (y_i > x_i)$$

More p-value calculation

- Individual feature probabilities calculated empirically.
- Example:

Vector	a-b	a-d	a-e	a-f	b-c	b-d	c-e	c-f	d-f
G_1	2	0	3	0	1	1	0	0	0
G_2	4	0	0	0	2	1	0	0	1
G_3	3	0	0	0	1	2	1	1	0
G_4	0	3	0	3	0	0	0	0	2

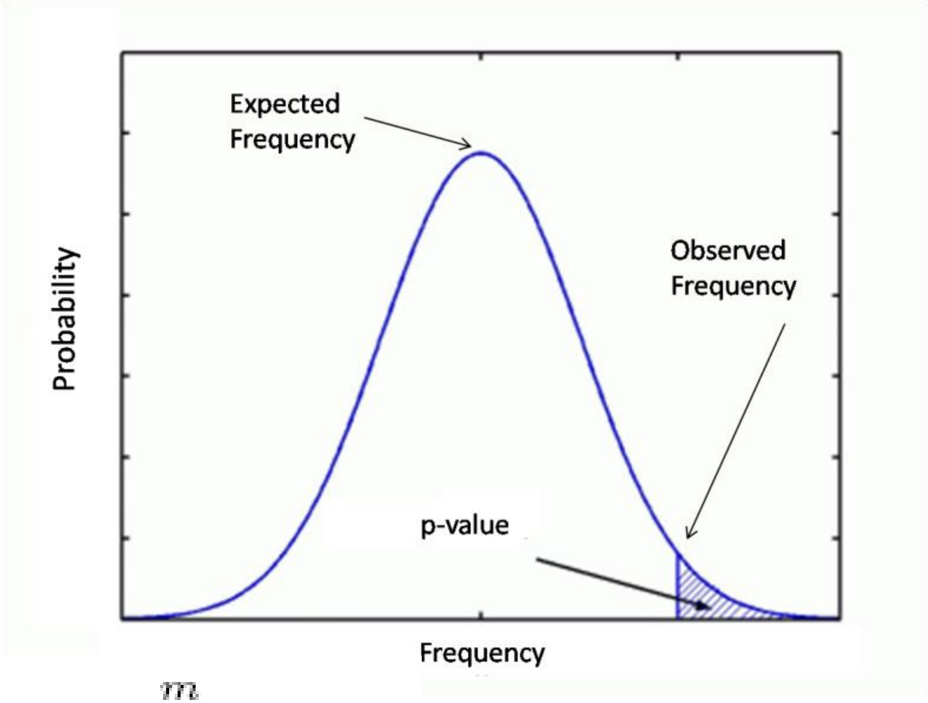
- P(a-b≥2)=3/4
- P(a-e≥ 1)=1/4
- P([2,0,0,0,1,1,0,0,0])= ¾* ¾* ¾ = 27/64

Probability Distribution of $\underline{X}$

- The distribution can be modeled as a Binomial Distribution
  - $P(\underline{x};\mu) = \binom{m}{\mu} P(\underline{x})^{\mu}(1 - P(\underline{x}))^{m-\mu}$
  - m = number of vectors in database
  - μ = number of successes

- X occurring in a vector a “success”

P-value…



- $$p\text{-}value(x, \mu_0) = \sum_{i=\mu_0}^{m} P(\underline{x}; i)$$
- μ_0 = observed frequency

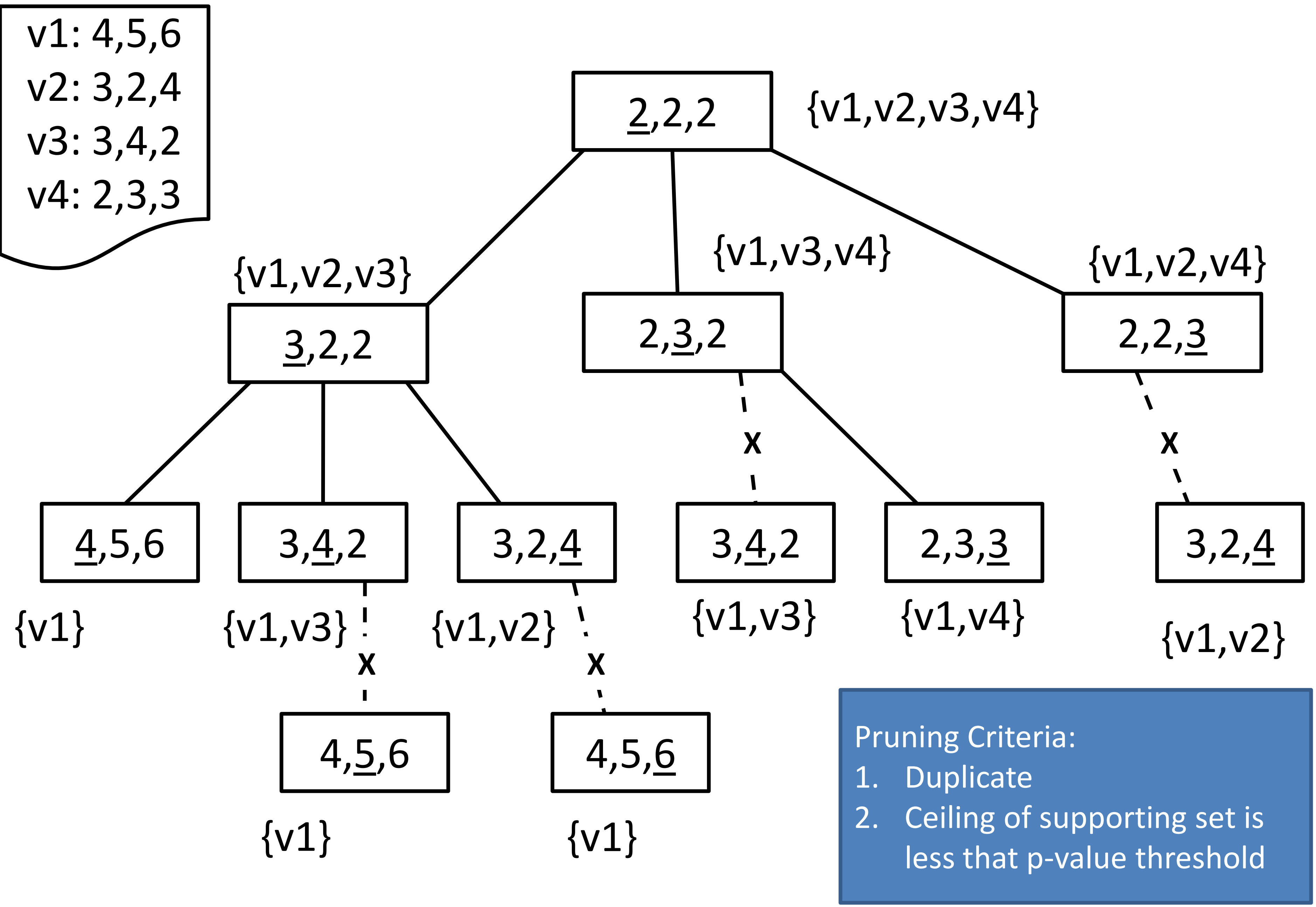
Monotonicity properties of p-value

- If $\underline{X}$ is a sub-feature vector of $\underline{Y}$
  – p-value($\underline{X}$,s) ≥p-value($\underline{Y}$,s) for any support s

- For some support $s_1 \geq s_2$
  – p-value($\underline{X}$,s_1) ≤ p-value($\underline{X}$,s_2)

Mining Significant subgraphs

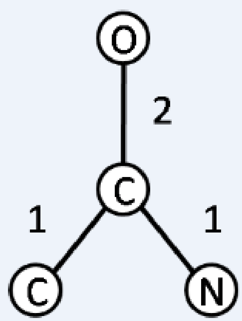
- What have we developed till now?
  - Vector representation of subgraphs
  - Significance of a subgraph using its vector representation
- Next Step?
  - Find all significant vectors


v1: 4,5,6
v2: 3,2,4
v3: 3,4,2
v4: 2,3,3
2,2,2
{v1,v2,v3,v4}
{v1,v2,v3}
3,2,2
{v1,v3,v4}
2,3,2
{v1,v2,v4}
2,2,3
x
x
4,5,6
3,4,2
3,2,4
3,4,2
2,3,3
3,2,4
{v1}
{v1,v3}
{v1,v2}
{v1,v3}
{v1,v4}
{v1,v2}
x
x
4,5,6
4,5,6
{v1}
{v1}
Pruning Criteria:
1. Duplicate
2. Ceiling of supporting set is less that p-value threshold

Definitions

- Vector $\underline{X}$ occurs in graph G
  - $\underline{X} \leq \underline{h}_i, \underline{h}_i \in G$
  - *Ex: [3,1,2] occurs in G, [3, 3, 3]* does not.

Graphical Representation

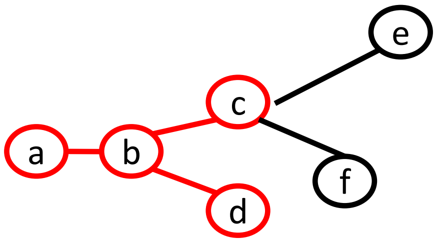


Random Walk Results

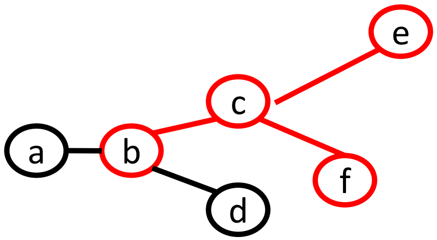
ID	Starting Atom	O-2-C	C-1-C	C-1-N
h_1	O	4	2	2
h2	C	2	3	3
h3	C	2	4	2
h4	N	2	2	4

Definitions..

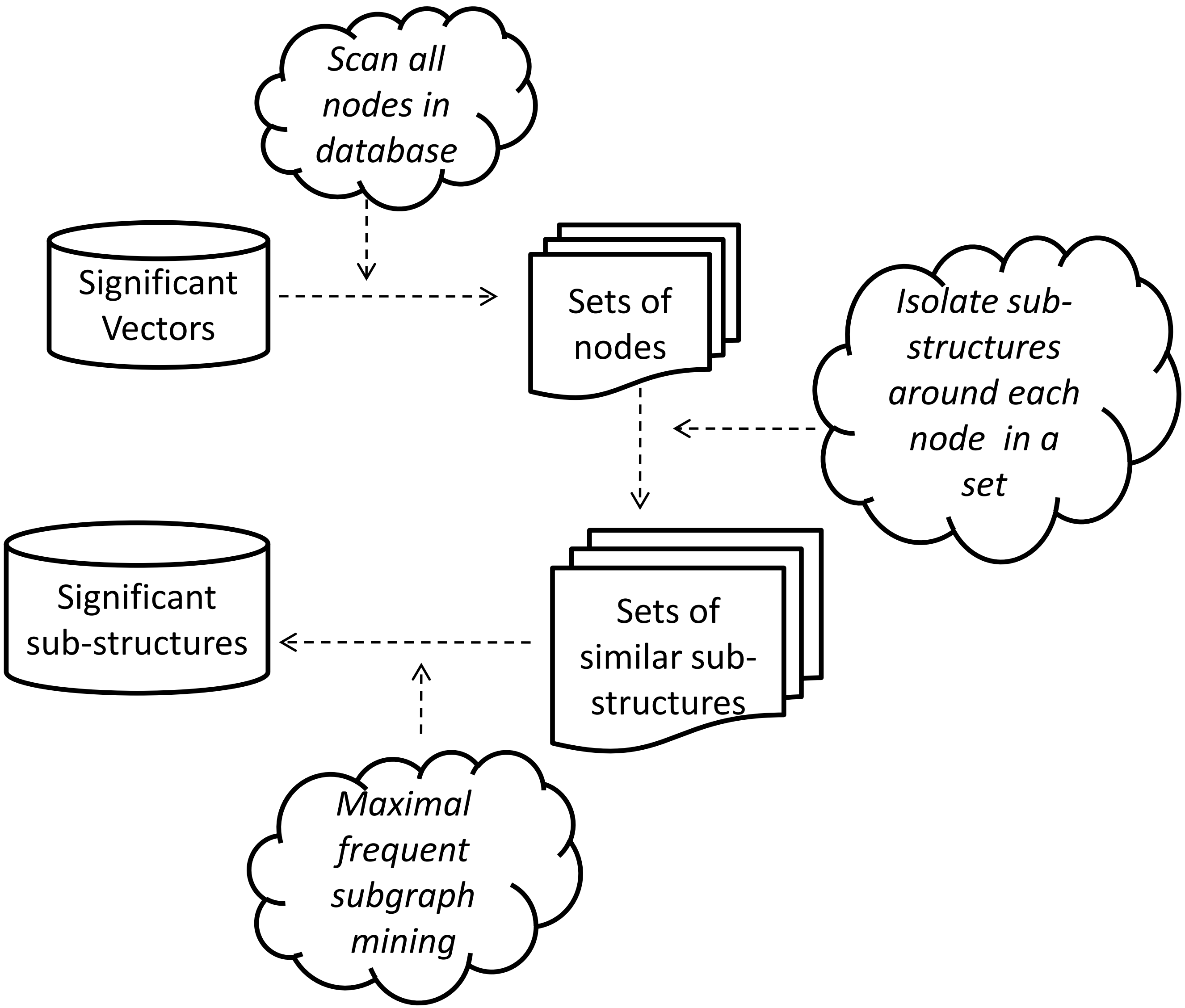
- Cut-off/Isolate structure around node *n* in Graph G within radius *r*
  - *Ex:* around **b** within radius **1**



  - *Ex:* around **f** within radius **2**



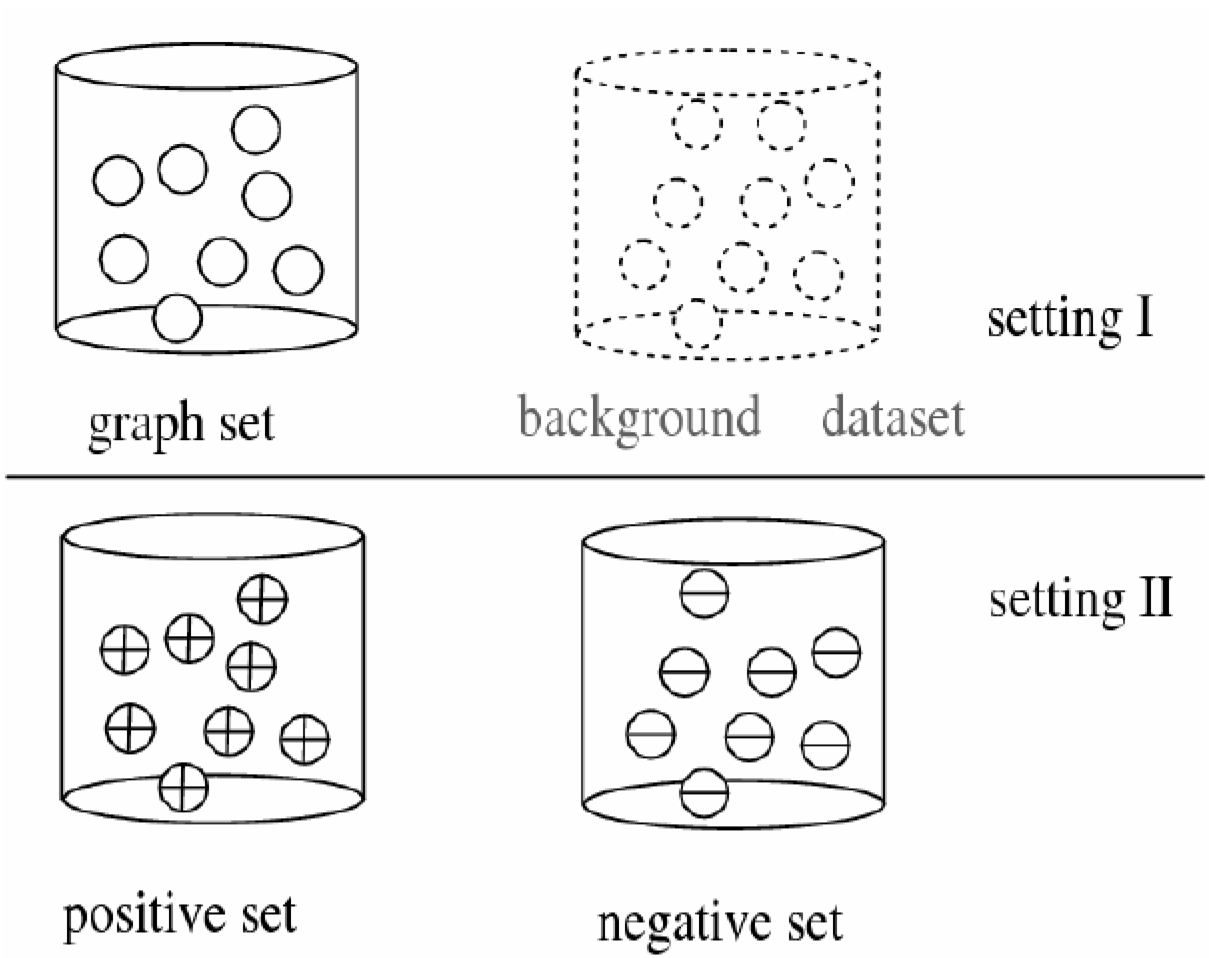
Mapping significant vectors to significant subgraphs



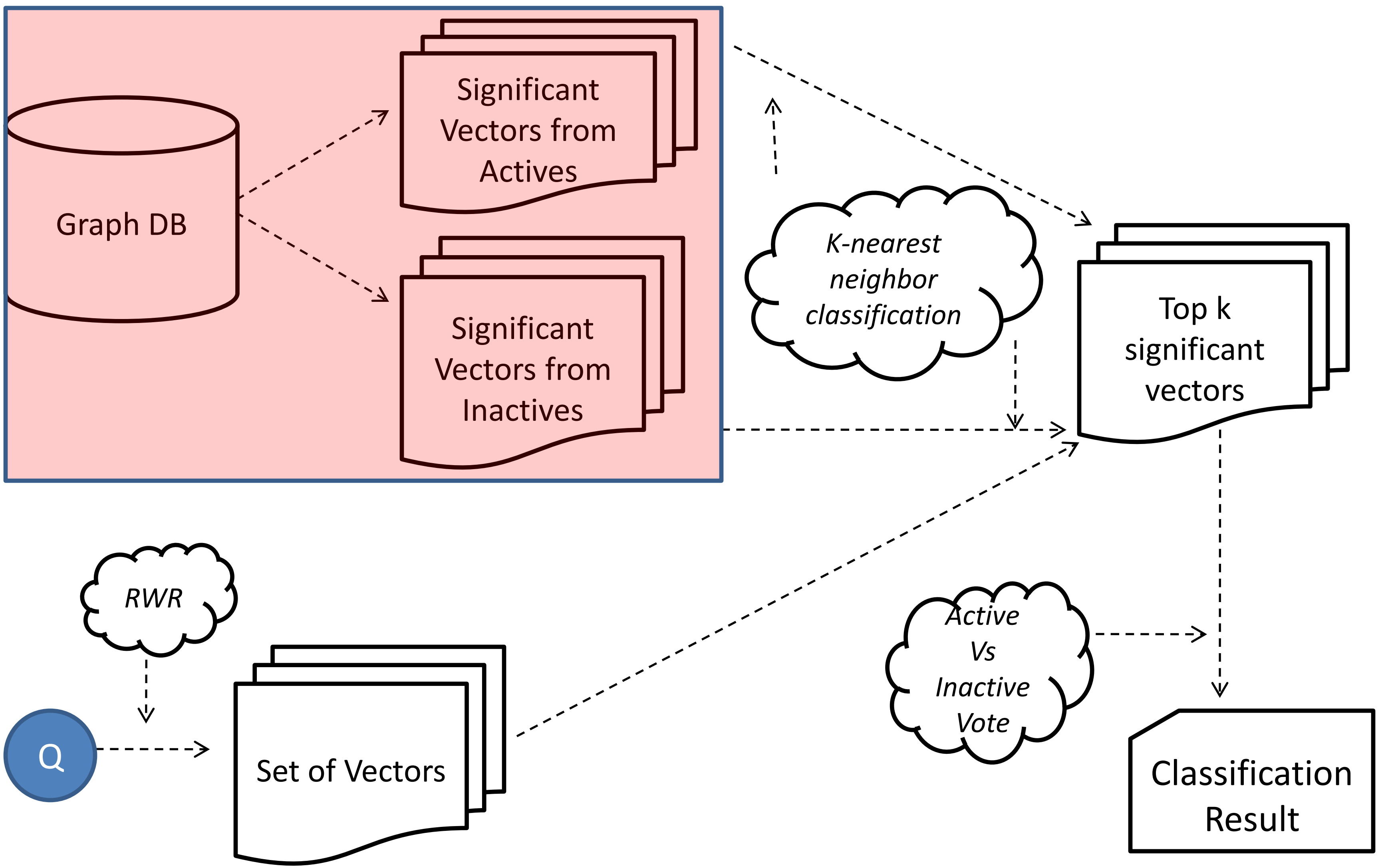
Application of significant subgraphs

- Over-represented molecular sub-structures
- Graph Classification
  - Significant subgraphs are more efficient than frequent subgraphs

Graph Setting



Classification Flowchart

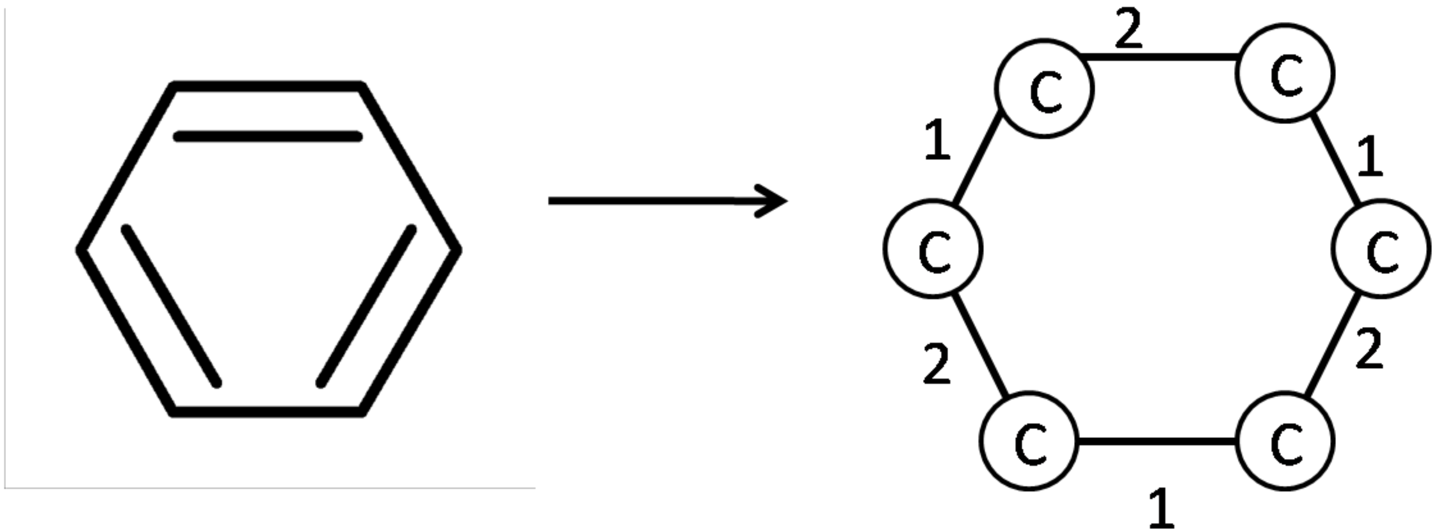


Experimental Results: Datasets

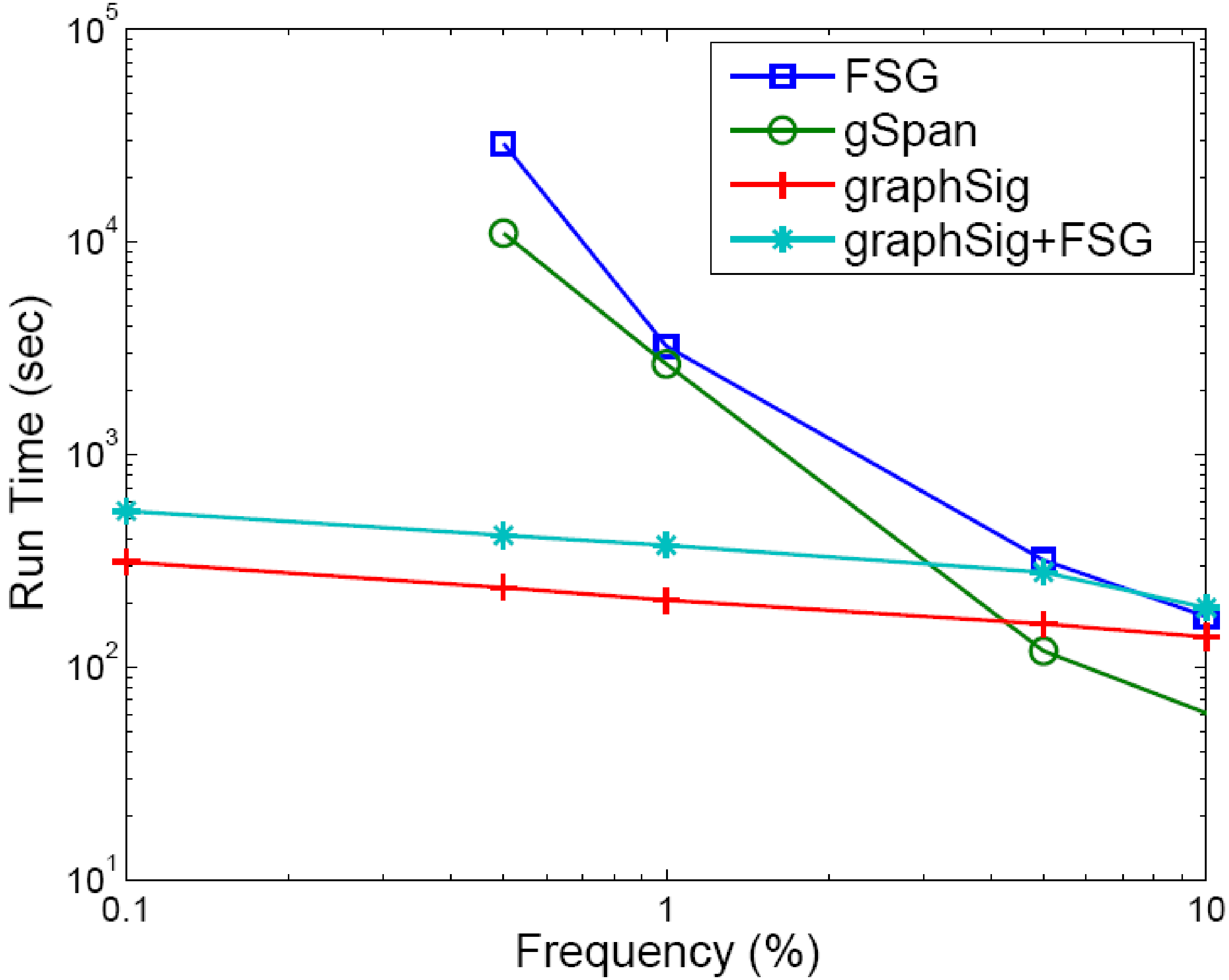
- AIDS dataset
- Cancer Datasets

Name	Size	Description
MCF-7	28972	Breast
MOLT-4	41810	Leukemia
NCI-H23	42164	Non-Small Cell Lung
OVCAR-8	42386	Ovarian
P388	46440	Leukemia
PC-3	28679	Prostate
SF-295	40350	Central Nervous System
SN12C	41855	Renal
SW-620	42405	Colon
UACC-257	41864	Melanoma
Yeast	83933	Yeast anticancer

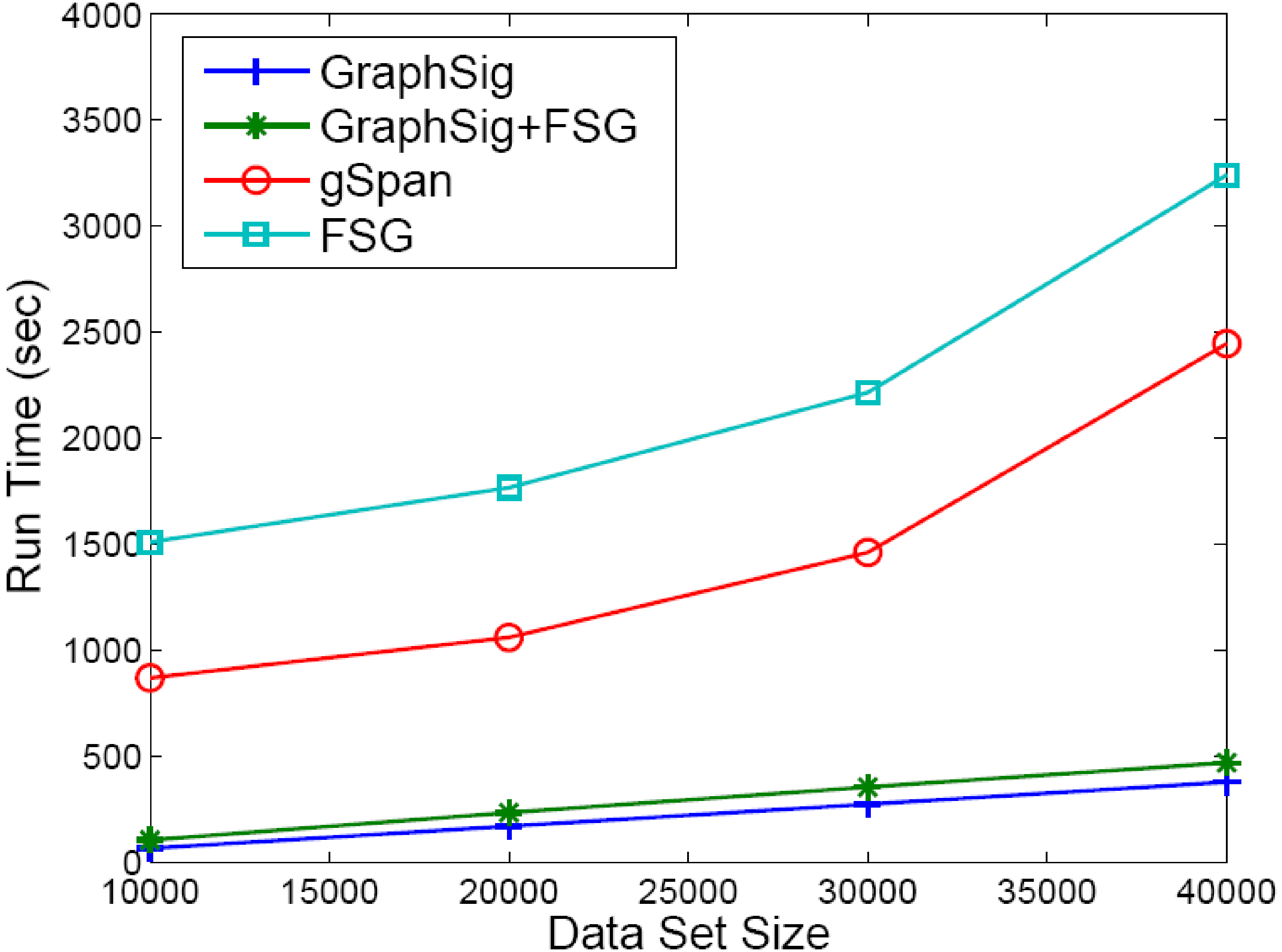
Representing molecules as graphs



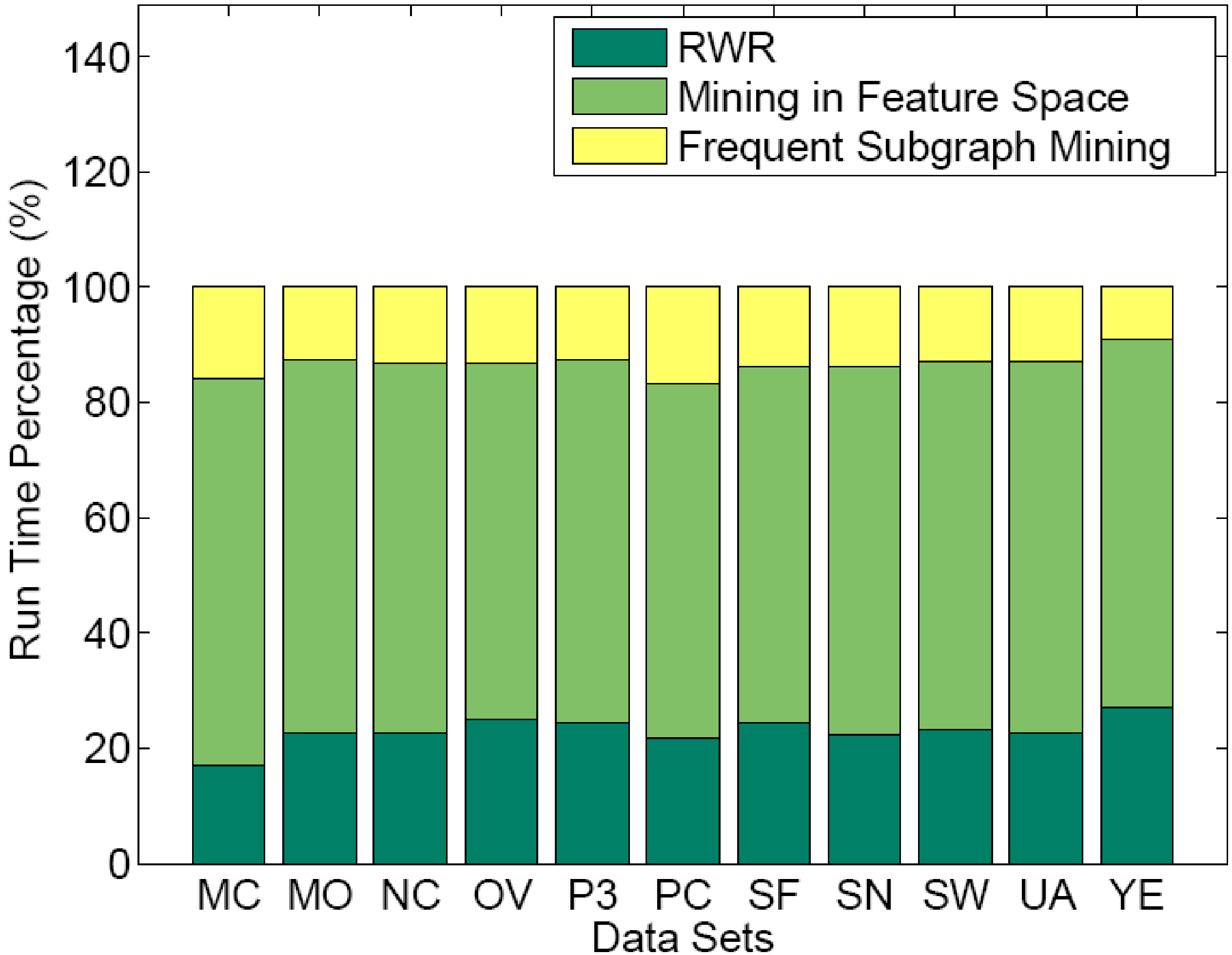
Time Vs. Frequency



Time vs DB size

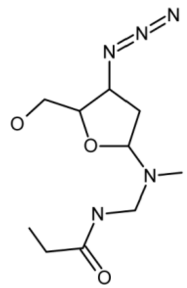


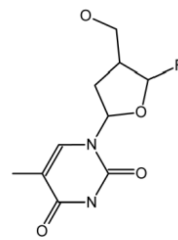
Profiling of Computation Cost



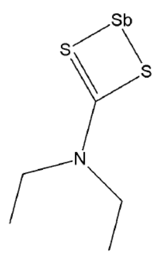
Quality of Patterns

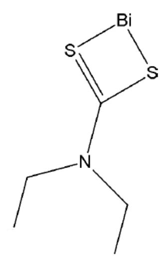
- Subgraphs mined from AIDS database





- Subgraphs mined from molecules active against Leukemia

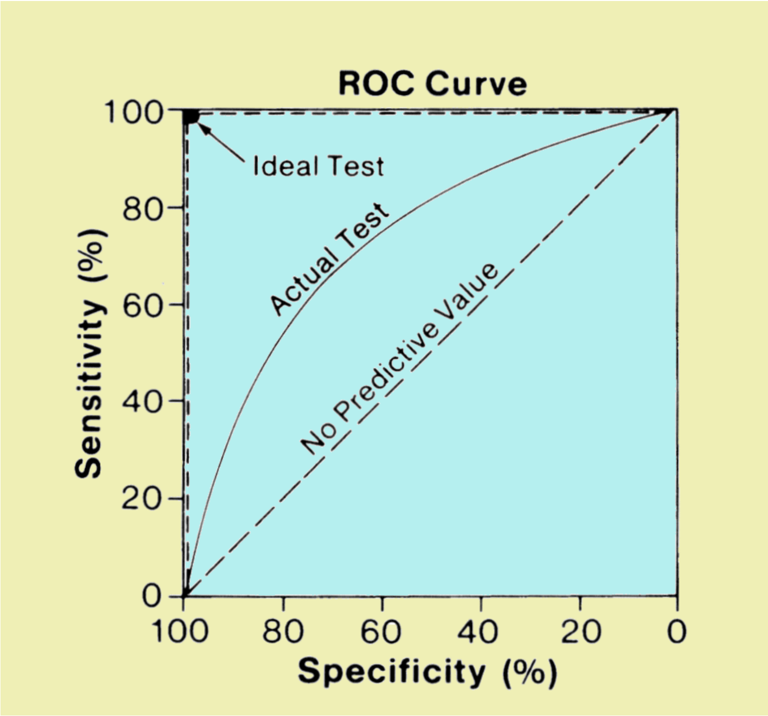




  – Sb and Bi are found at a frequency below 1%
  – Current techniques unable to scale to such low frequencies

Classification

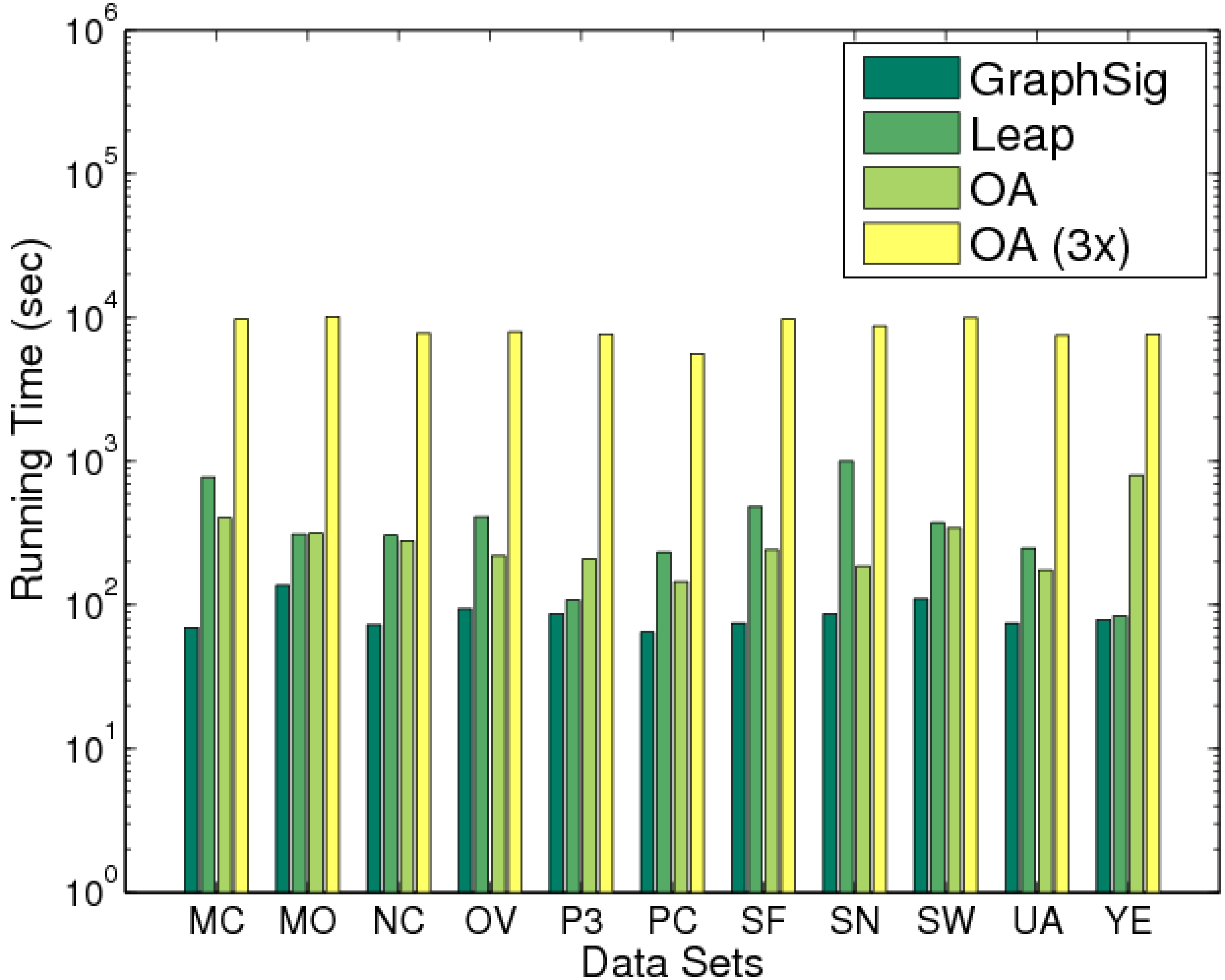
- Performance Measure: Area under ROC Curve (AUC)
- AUC is between 0 and 1.
- Higher the AUC better the performance.



AUC Comparison

Dataset	**OA Kernel**	**Leap**	**GraphSig**
MCF-7	0.68 ± 0.12	0.76 ± 0.04	**0.77 ± 0.02**
MOLT-4	0.65 ± 0.06	0.72 ± 0.06	**0.74 ± 0.02**
NCI-H23	0.79 ± 0.08	0.79 ± 0.05	**0.80 ± 0.02**
OVCAR-8	0.67 ± 0.04	0.78 ± 0.02	**0.79 ± 0.02**
P388	0.79 ± 0.07	**0.84 ± 0.03**	**0.84 ± 0.02**
PC-3	0.66 ± 0.09	**0.76 ± 0.04**	**0.76 ± 0.03**
SF-295	0.75 ± 0.11	0.77 ± 0.02	**0.80 ± 0.02**
SN12C	0.75 ± 0.08	**0.80 ± 0.02**	**0.80 ± 0.03**
SW-620	0.70 ± 0.02	0.76 ± 0.04	**0.77 ± 0.02**
UACC-257	0.65 ± 0.05	0.75 ± 0.03	**0.81 ± 0.02**
Yeast	0.64 ± 0.04	0.71 ± 0.02	**0.73 ± 0.04**
Average	0.702 ± 0.07	0.767 ± 0.03	**0.782 ± 0.02**

Running Time Comparison



Questions?