

Efficient Mining of Frequent Subgraphs in Graph Databases

Jun(Luke) Huan, Wei Wang, Jan Prins

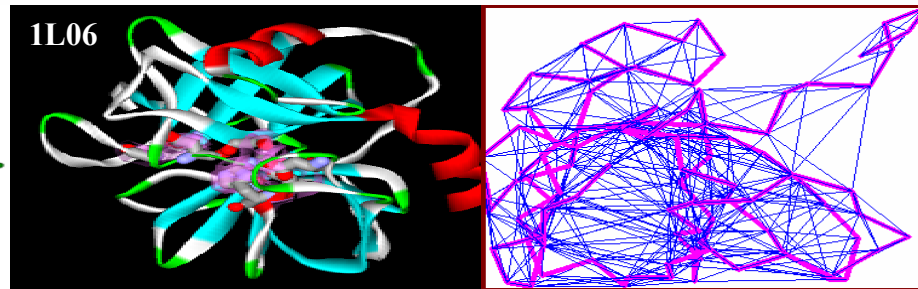
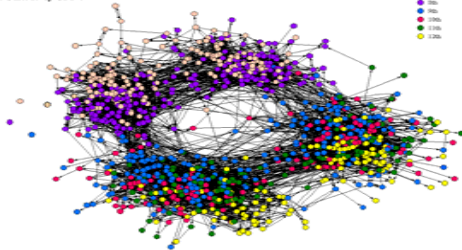
Nov 19, 2003



Presentation Overview

- Introduction
 - Finding recurring subgraphs from graph databases.
- CSM: Coherent Subgraph Mining Algorithm
 - Novel Data Structure: CAM tree
 - Mutual Information Selection
- Experimental Study
 - Classifying SCOP protein family

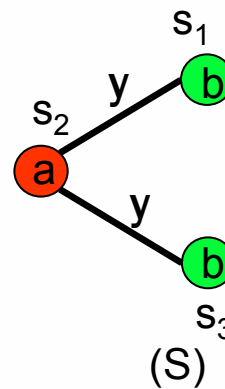
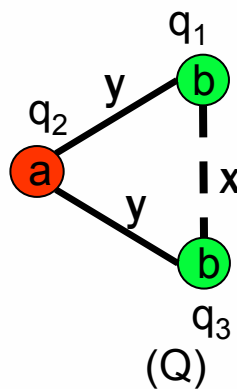
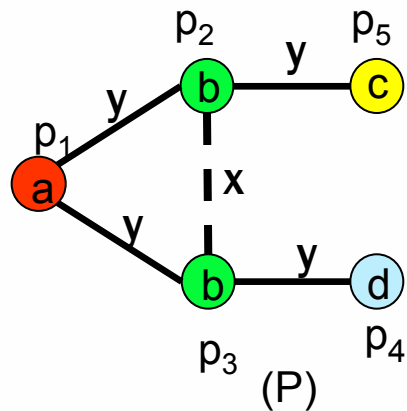
The Social Structure of "Countryside" School District
Points Colored by Grade





Labeled Graph

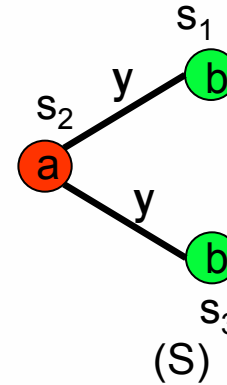
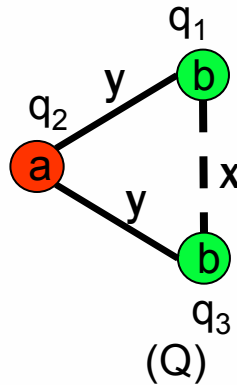
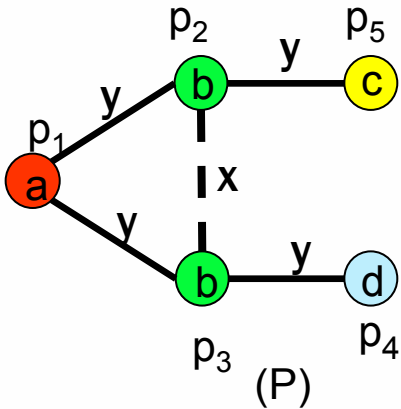
- We define a labeled graph G as a five element tuple $G = \{V, E, \Sigma_V, \Sigma_E, \delta\}$ where
 - V is the set of vertices of G ,
 - $E \subseteq V \times V$ is a set of undirected edges of G ,
 - Σ_V (Σ_E) are set of vertex (edge) labels,
 - δ is the labeling function: $V \rightarrow \Sigma_V$ and $E \rightarrow \Sigma_E$ that maps vertices and edges to their labels.





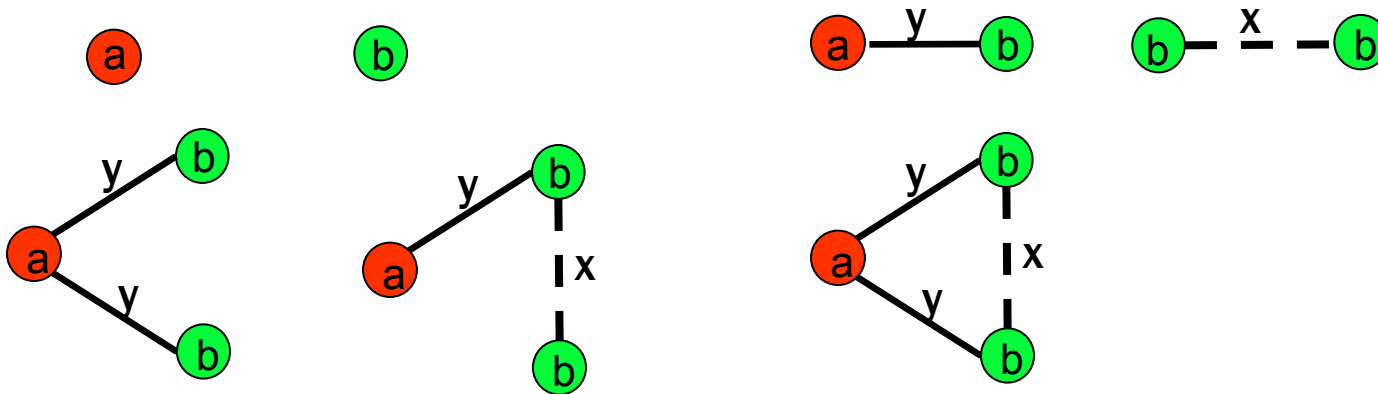
Frequent Subgraph Mining

Input: A set GD of labeled undirected graphs



$\sigma = 2/3$

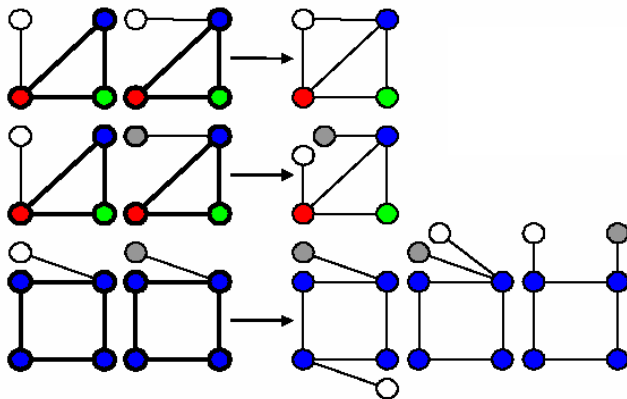
Output: All frequent subgraphs (w. r. t. σ) from GD .



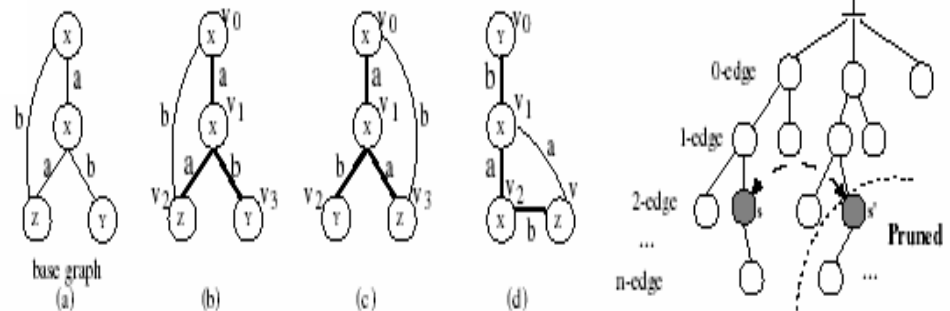


FFSM: Fast Frequent Subgraph Mining -- An Overview:

- How to solve graph isomorphism problem?
 - A Novel Graph Canonical Form: CAM
- How to solve subgraph isomorphism problem (NP-complete)?
 - Incrementally kept embeddings
- How to enumerate subgraphs:
 - An Efficient Data Structure: CAM Tree
 - Two Operations: CAM-join, CAM-extension.



FSG, M. Kuramochi & G. Karypis, ICDM'01

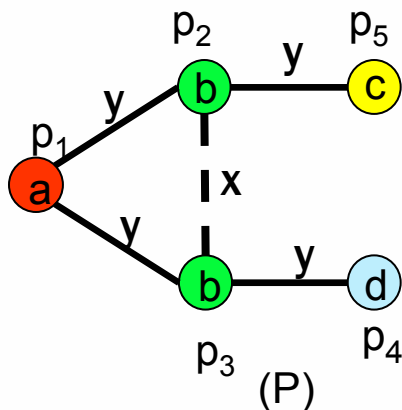


gSpan, X. Yan and J. Han, ICDM'02



Adjacency Matrix

- Every diagonal entry of adjacency matrix M corresponds to a distinct vertex in G and is filled with the label of this vertex.
- Every off-diagonal entry in the lower triangle part of M^1 corresponds to a pair of vertices in G and is filled with the label of the edge between the two vertices and zero if there is no edge.



a				
y	b			
y	x	b		
0	y	0	c	
0	0	y	0	d

M_1

a				
y	b			
y	x	b		
0	0	y	d	
0	y	0	0	c

M_2

b				
x	b			
y	0	d		
0	y	0	c	
y	y	0	0	a

M_3

¹for an undirected graph, the upper triangle is always a mirror of the lower triangle



Code

- A Code of $n \times n$ adjacency matrix M is defined as sequence of lower triangular entries (including the diagonal entries) in the order:

$$M_{1,1} M_{2,1} M_{2,2} \cdots M_{n,1} M_{n,2} \cdots M_{n,n-1} M_{n,n}$$

a				
y	b			
y	x	b		
0	y	0	c	
0	0	y	0	d

M_1

a				
y	b			
y	x	b		
0	0	y	d	
0	y	0	0	c

M_2

b				
x	b			
y	0	d		
0	y	0	c	
y	y	0	0	a

M_3

Code(M_1): aybyxb0y0c00y0d >

Code(M_2): aybyxb00yd0y00c >

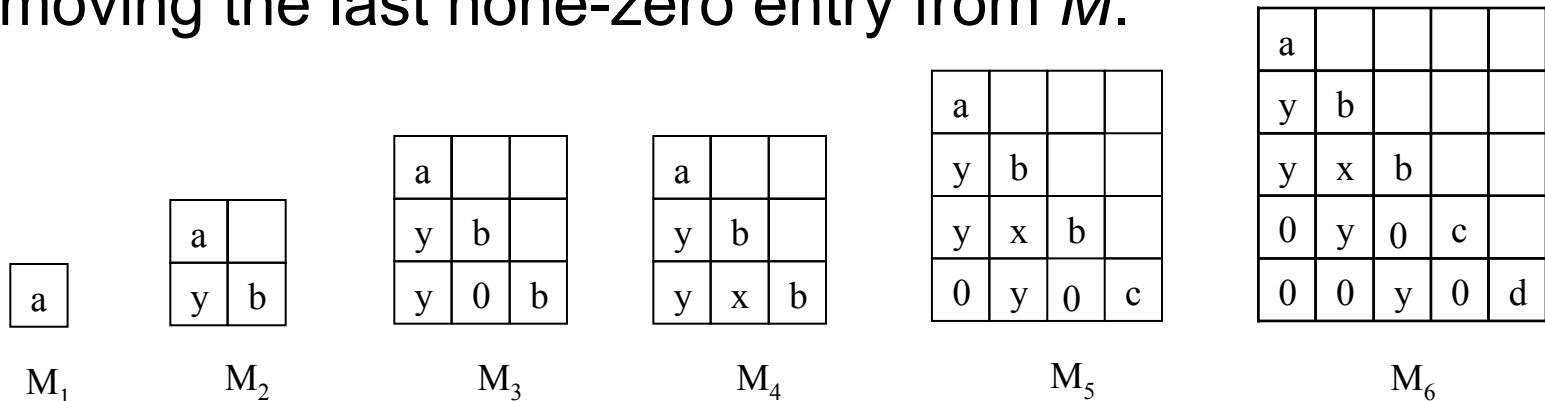
Code(M_3): bxby0d0y0cyy00a

- The Canonical Adjacency Matrix is the one produces the maximal code, using lexicographic order.



MP Submatrix

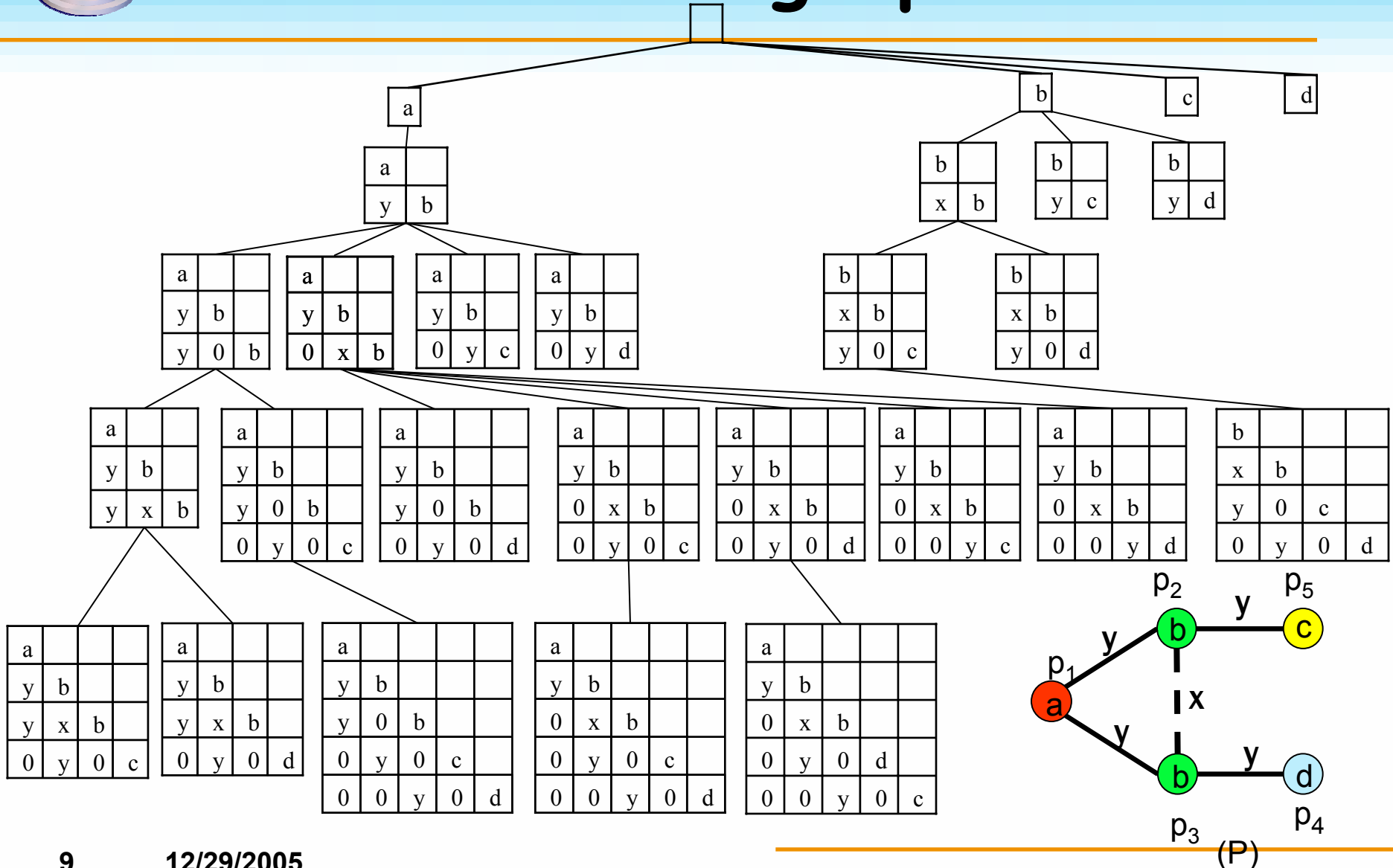
- For an $m \times m$ matrix A , an $n \times n$ matrix B is A 's maximal proper submatrix (MP Submatrix), iff N is obtained by removing the last none-zero entry from M .



- We define a CAM is connected iff the corresponding graph is connected.
- **Theorem I:** A CAM's MP submatrix is CAM
- **Theorem II:** A connected CAM's MP submatrix is connected



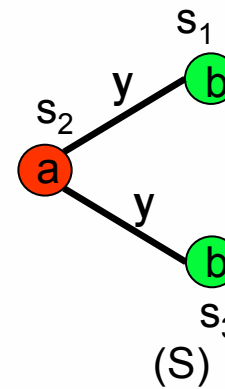
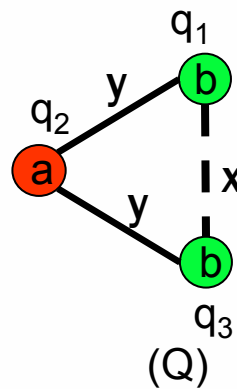
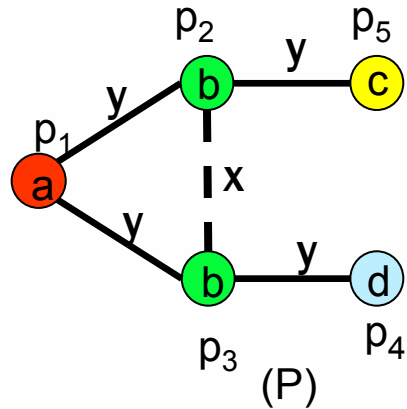
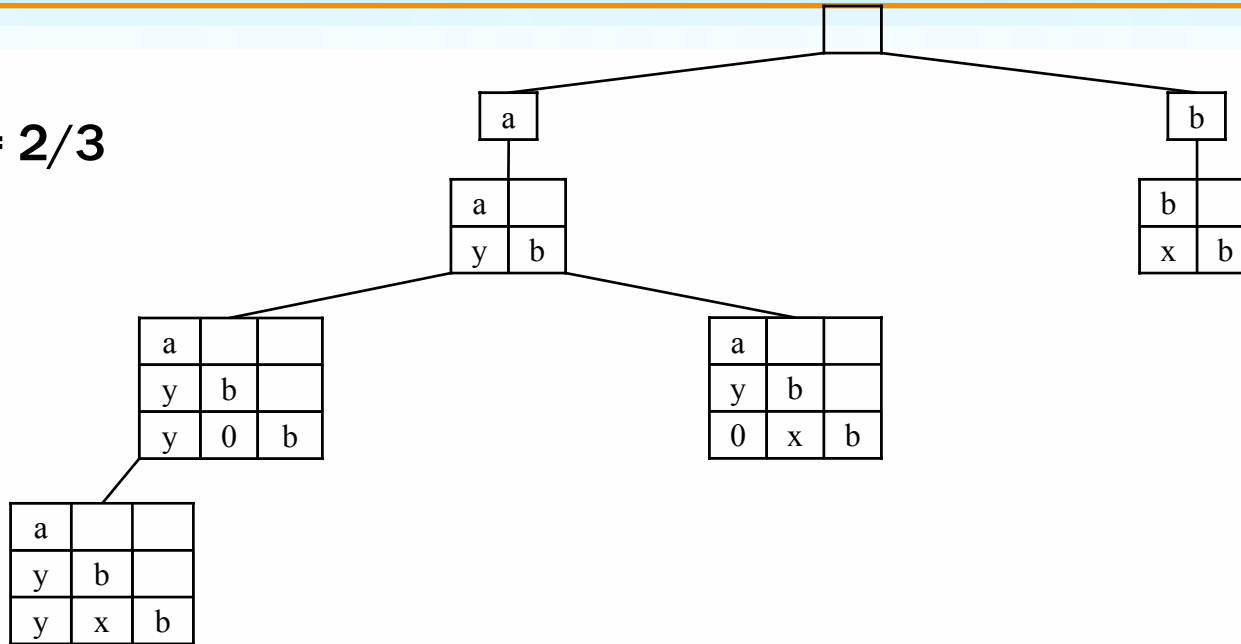
CAM Tree: Subgraphs





CAM Tree: Frequent Subgraphs

$\sigma = 2/3$





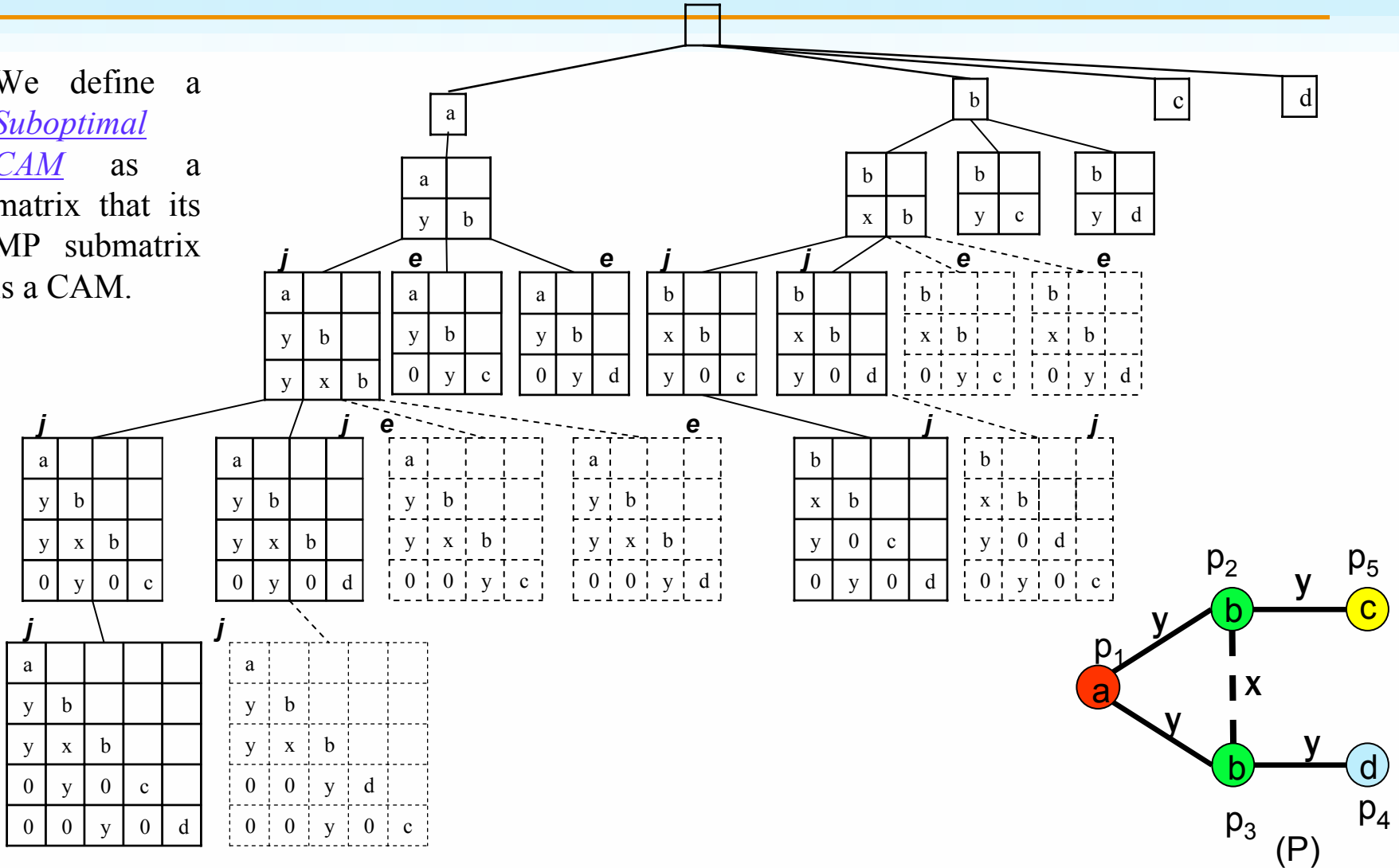
How to Enumerate Nodes in a CAM Tree?

- Two operations to explore CAM tree:
 - CAM-Join
 - CAM-Extension
- Augmenting CAM tree with Suboptimal CAMs
- Objectives:
 - none false dismissal
 - no redundancy
- Plus: We want to this **efficiently!**



Suboptimal Tree

We define a *Suboptimal CAM* as a matrix that its MP submatrix is a CAM.

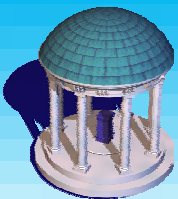




Summary

■ Theorem:

For a graph G , let C_{K-1} (C_K) be set of the suboptimal CAMs of all the size $(K-1)$ (K) subgraphs of G ($K \geq 2$). Every member of set C_K can be enumerated unambiguously either by **joining** two members of set C_{K-1} or by **extending** a member in C_{K-1} .

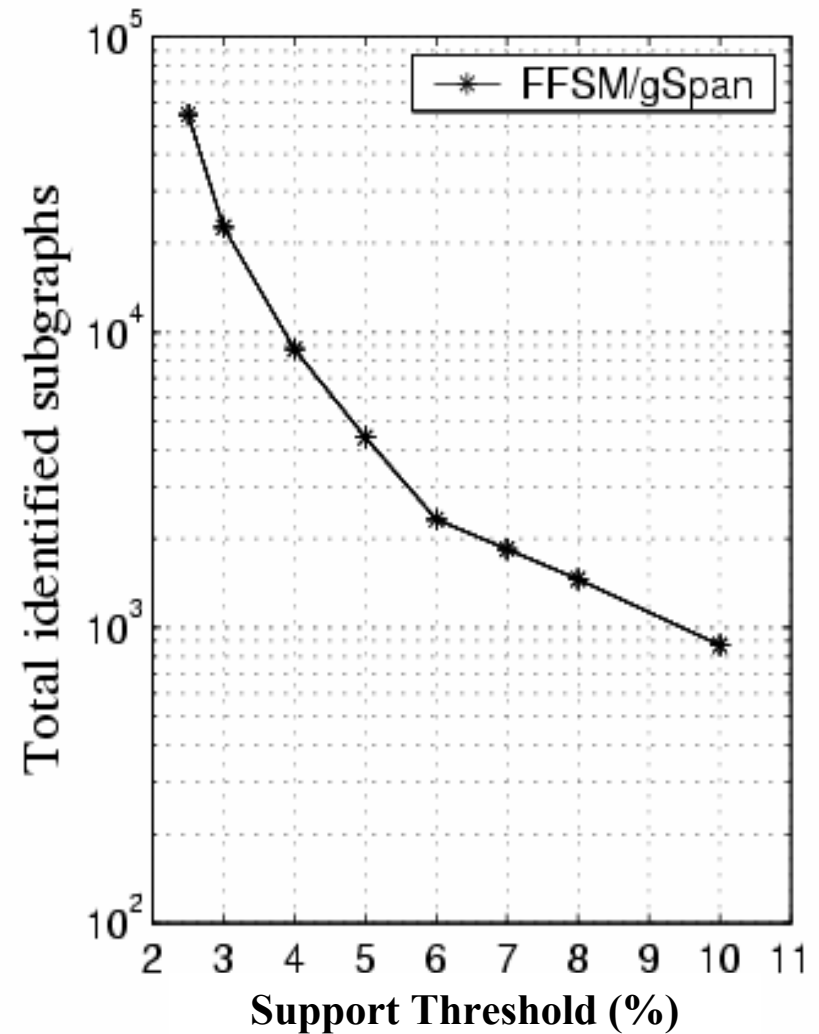
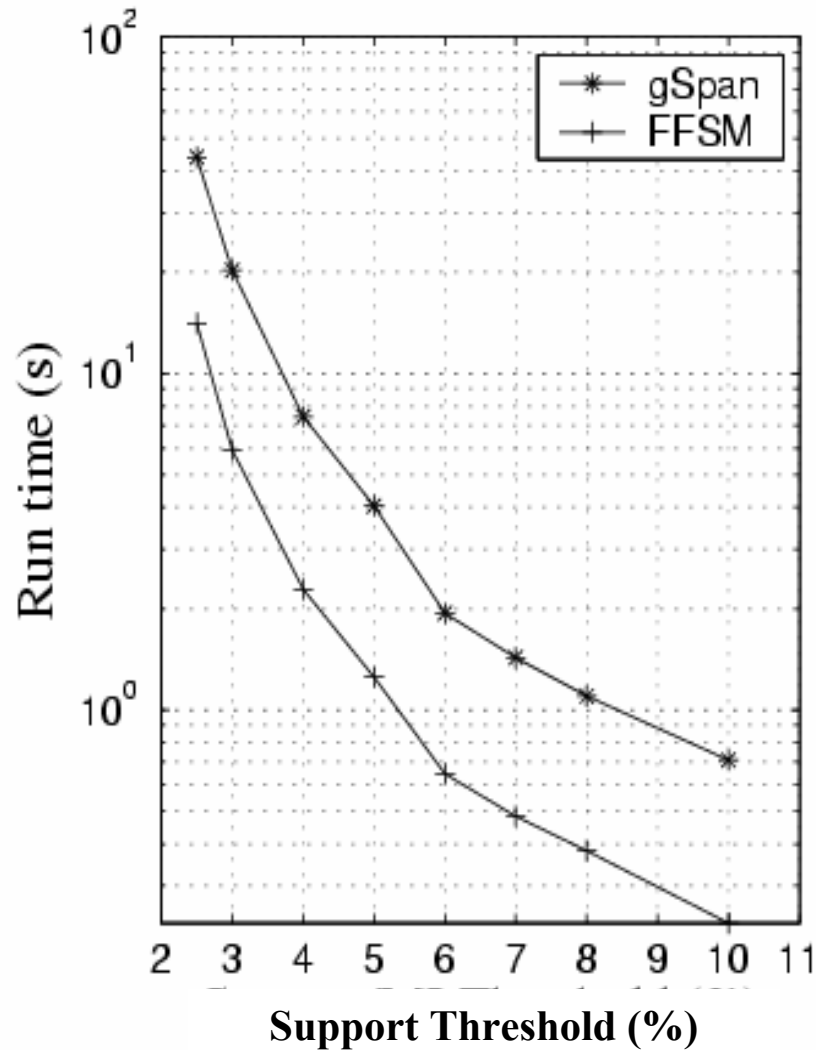


Experimental Study

- Predictive Toxicology Evaluation Competition (PTE)
 - Contains: 337 compounds
 - Each graph contains 27 nodes and 27 edges on average
- NIH DTP Anti-Viral Screen Test (DTP CA/CM)
 - Chemicals are classified to be Confirmed Active (CA), Confirmed Moderate Active (CM) and Confirmed Inactive (CI).
 - We formed a dataset contains CA (423) and CM (1083).
 - Each graph contains 25 nodes and 27 edges on average

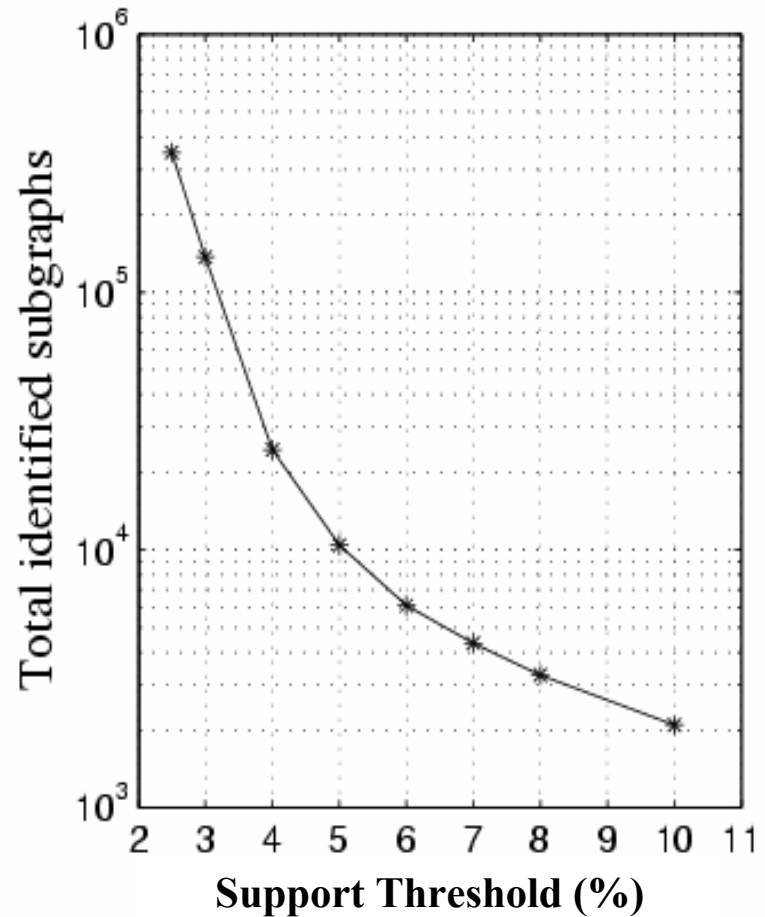
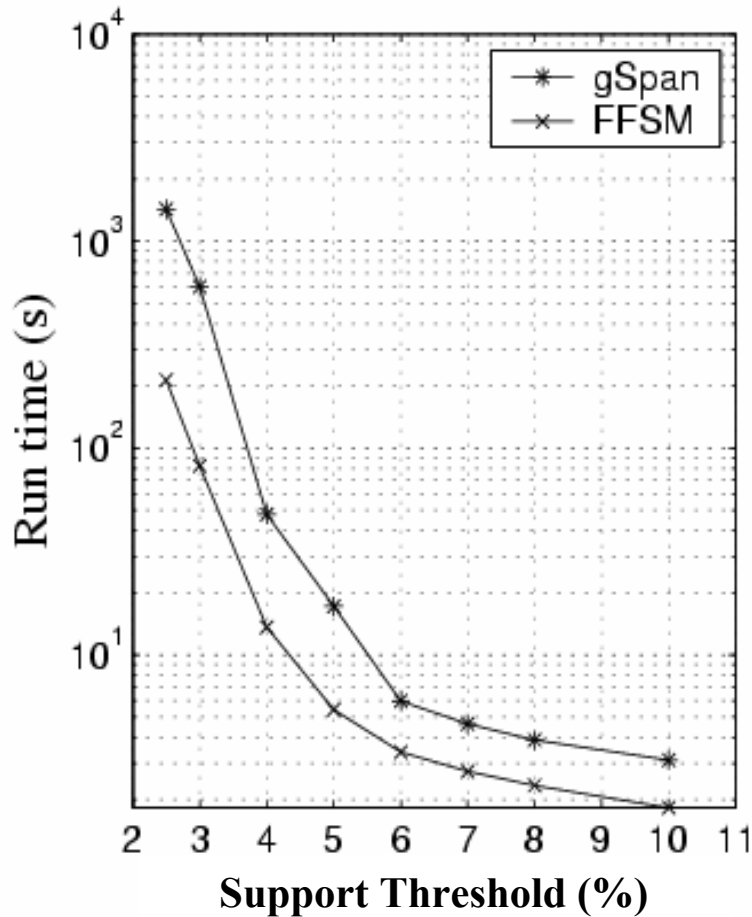


Performance (PTE)





Performance (DTP CACM)





Thank You



For More Information

Dr. Wei Wang

Department of Computer Science,
University of North Carolina at Chapel Hill,
Chapel Hill, NC 27510

Email: weiwang@cs.unc.edu

<http://www.cs.unc.edu/~weiwang>