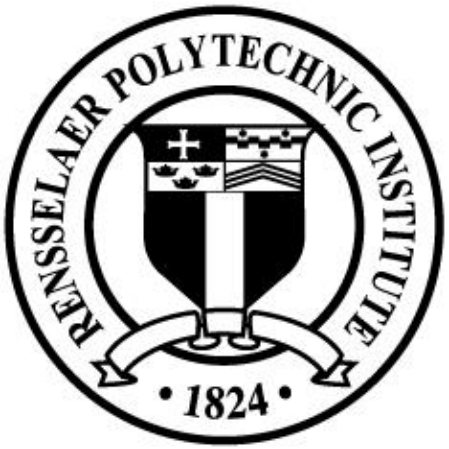
Pattern Mining: Past, Present & Future

Mohammed J. Zaki
Rensselaer Polytechnic Institute (RPI)
Troy NY

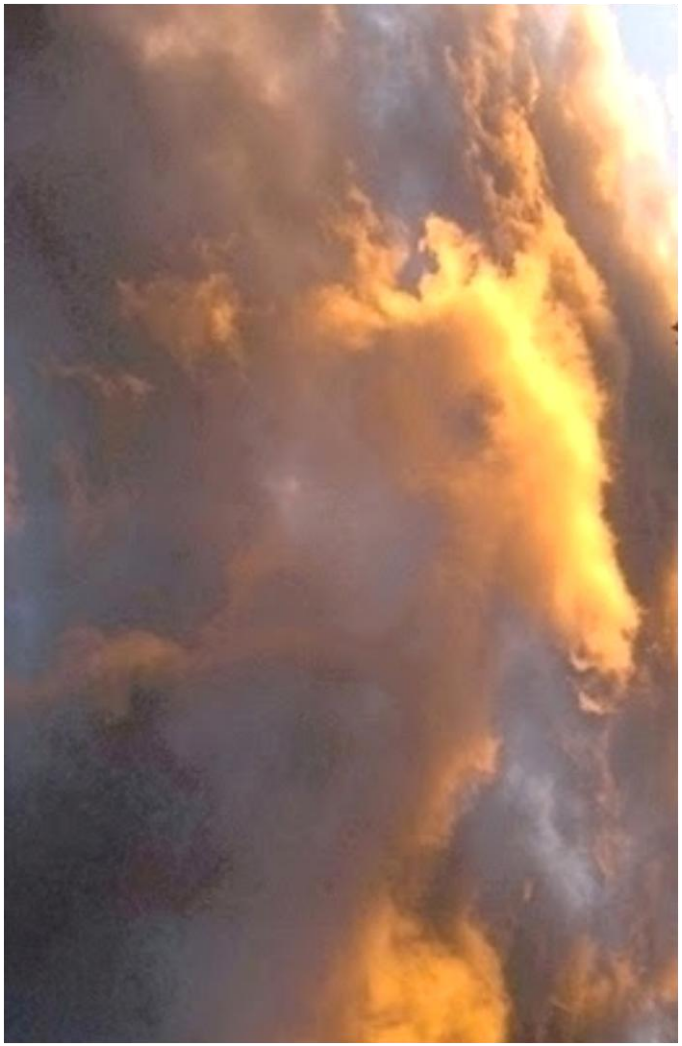


An outline

- A personal (biased) perspective
  - Not comprehensive; usual disclaimers!
  - Data mining perspective
    - Completely ignores parallel developments in combinatorial pattern matching, bioinformatics, network science
- Bit of history
- Some applications
- A bit of the present & future …

Are humans inherently good at pattern mining? Is there a pattern?

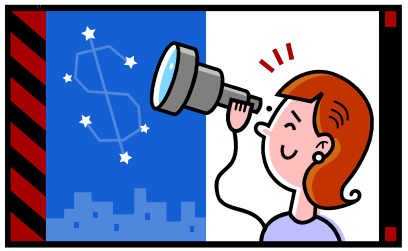
Pattern or Illusion?



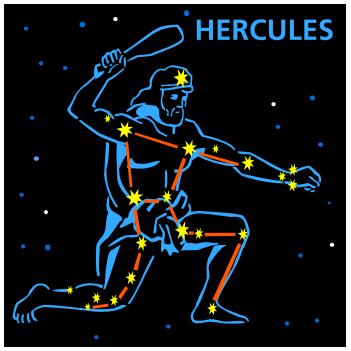
Pattern or Illusion?

Pattern Mining: Distant Past

- Humans have always been doing pattern mining (?)
  - Observing & predicting nature
    - The terra-firma (flora, fauna)
    - The heavens (climate, navigation)



- Humans generally good at pattern recognition
  - Illusion vs. pattern?
  - We see what we want to see! (bias)
  - Restricted to the "natural" dimensionality: 3D



Pattern vs. Chance

Dog is not the pattern; the black patches are!
But is that an interesting pattern?

What is a pattern?

- Repetitiveness
  - Basically depends on counting
- Interestingness
  - Avoid trivial patterns
- Chance occurrences
  - Use statistical tests to weed these out
- Rarity
  - Leave to anomaly detection

Pattern Mining: The Past

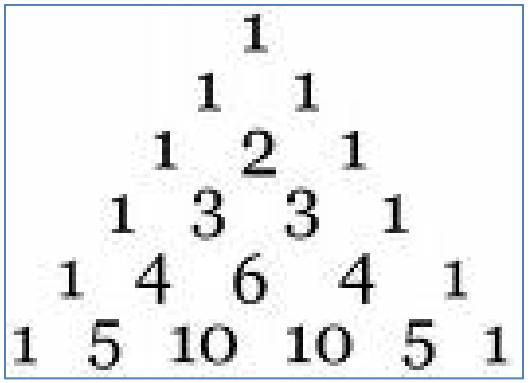
- In the beginning it was market baskets,

and it was all diapers & beer …

- Or was it?

It's all about sets

- Circa 1993: Agrawal, Imielinski & Swami introduce the concept of association rules & frequent itemsets for market basket data
- 1994: The classic method "Apriori" is proposed by Agrawal and Srikant (AS)
- 1994: Mannila, Toivonen, Verkamo (MTV) propose levelwise method
- 1995: AS and MTV combine their independent methods
- And a revolution is born!



But there is more…

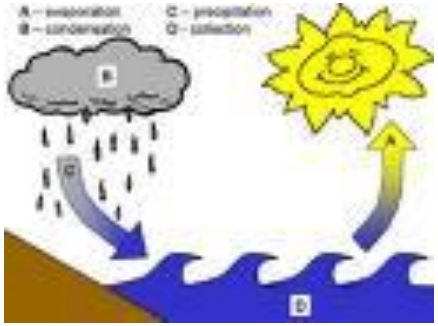
- Circa 1982: Wille invents formal concept analysis (FCA)
- Circa 1988: Luxenburger introduces the notion of "partial implications" which are essentially association rules without the frequent part
- Circa 1998: Marriage of Association Rules and FCA: frequent closed itemsets are born
  - Independently by
    - Zaki & Ogihara (DMKD'98)
    - Pasquier, Bastide, Taouil, Lakhal (BDA'98; in Hammamet!)

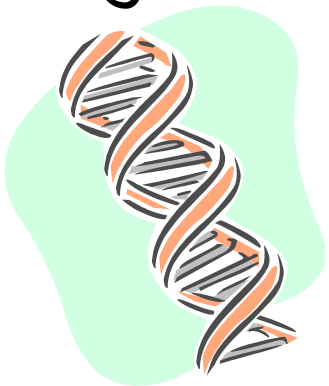
Other noteworthy events

- Alternative Algorithmic Paradigms
  - Vertical tidsets (ECLAT) by Zaki et al, 1997
  - FP-Growth by Han et al, 2000
- Maximal itemsets
  - Bayardo, 1998 (also mentioned in Zaki et al, 1997)
- Summarization
  - Closed itemsets (ZO & PBTL, 1998)
  - Free sets (Boulicaut, Bykowski, Rigotti, 2000)
  - Minimal Generators (Bastide, Taouil, Pasquier, Stumme, Lakhal, 2000)
  - Non-derivable itemsets (Calders & Goethals, 2002)
  - Active area of research (e.g. S. Ben Yahia, EGC'10)

What about sequences?

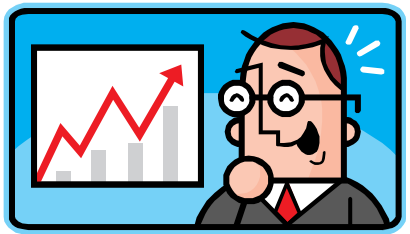
- 1995: Agrawal & Srikant propose sequential patterns
  - Notion of frequent sequences
  - Levelwise method like apriori
- String matching & sequence analysis has a much older history in combinatorial pattern matching & bioinformatics





Sequence Mining

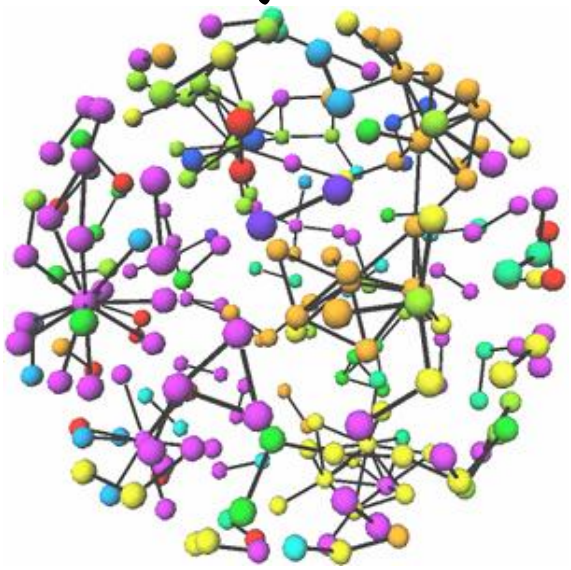
- Major paradigms
  - Levelwise: AS'95
  - Episode Mining: Mannila et al, 1995
  - Vertical (SPADE): Zaki 1998
  - Projection-based (prefixSPAN): Pei et al, 2001
- Summarization
  - Closed sequences: Yan et al, 2003



On to trees

- Induced trees: Wang & Liu, 1998
- Embedded trees: Zaki, 2002 (rightmost extension)
  - Similar candidate generation in Asai et al, 2002
- Maximal & closed trees:
  - Chi et al, 2004

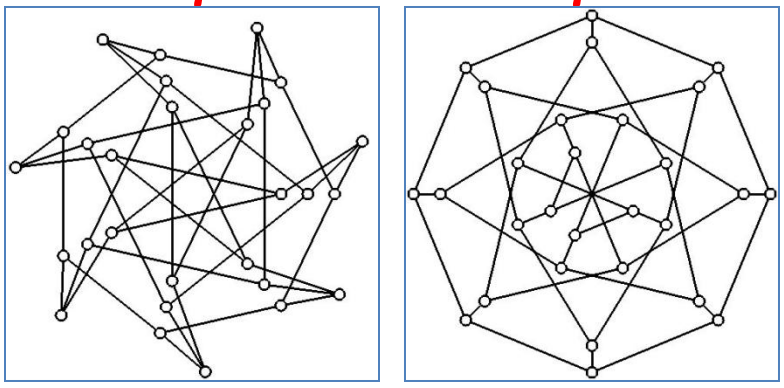
And then there are Graphs



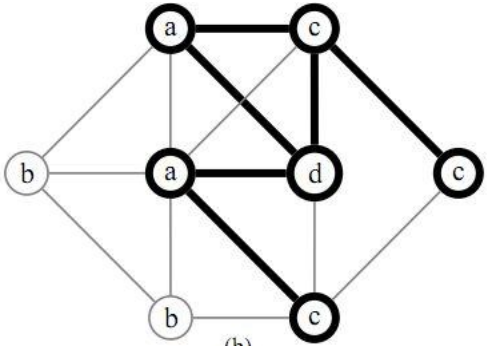
- Circa 1994-1995:
  - Heuristic Search
    - Cook & Holder, 1994
    - Yoshida & Motoda, 1995
- Frequent Subgraphs
  - AGM: Inokuchi, Washio, Motoda, 2000 (levelwise)
  - FGM: Kuramochi & Karypis, 2001 (levelwise)
  - gSpan: Yan & Han, 2002 (rightmost extension)
  - FSM: Huan et al, 2003 (canonical matrices)
  - Closed & Maximal graphs: Yan & Han, 2003; Huan et al, 2004, respv.

Taming of the Morphs

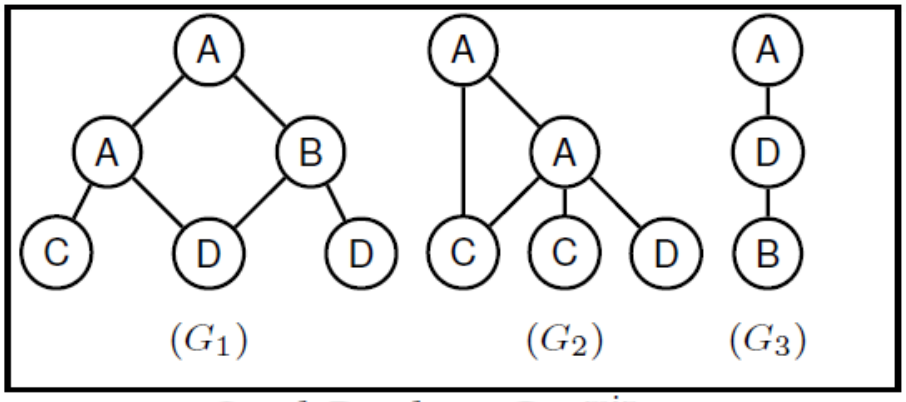
- Challenge of isomorphisms
- How to detect duplicates?
  - Graph Isomorphism



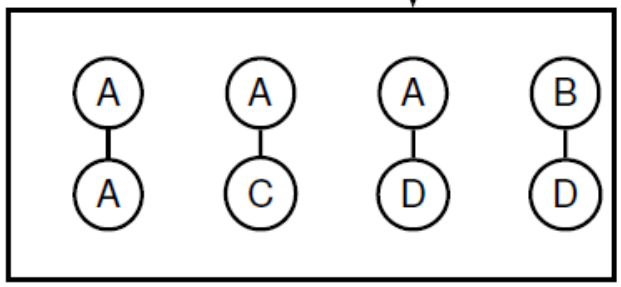
- How to count occurrences?
  - Subgraph Isomorphism



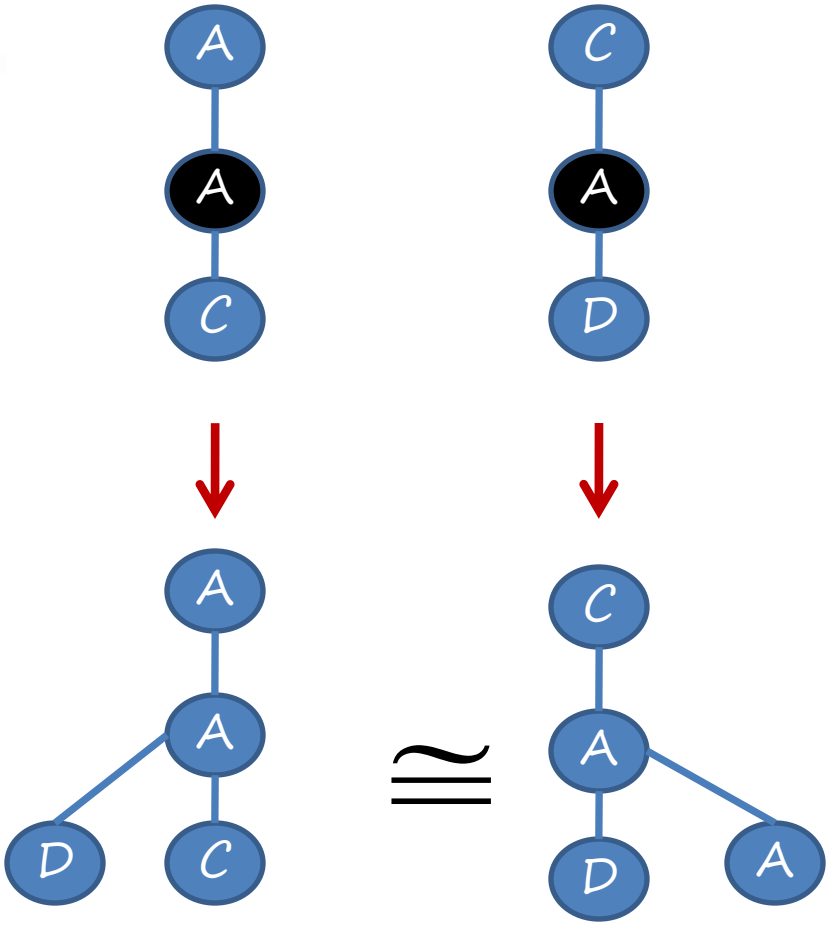
Candidate Generation



Graph Database, $\mathcal{D}, \pi^{\min} = 2$

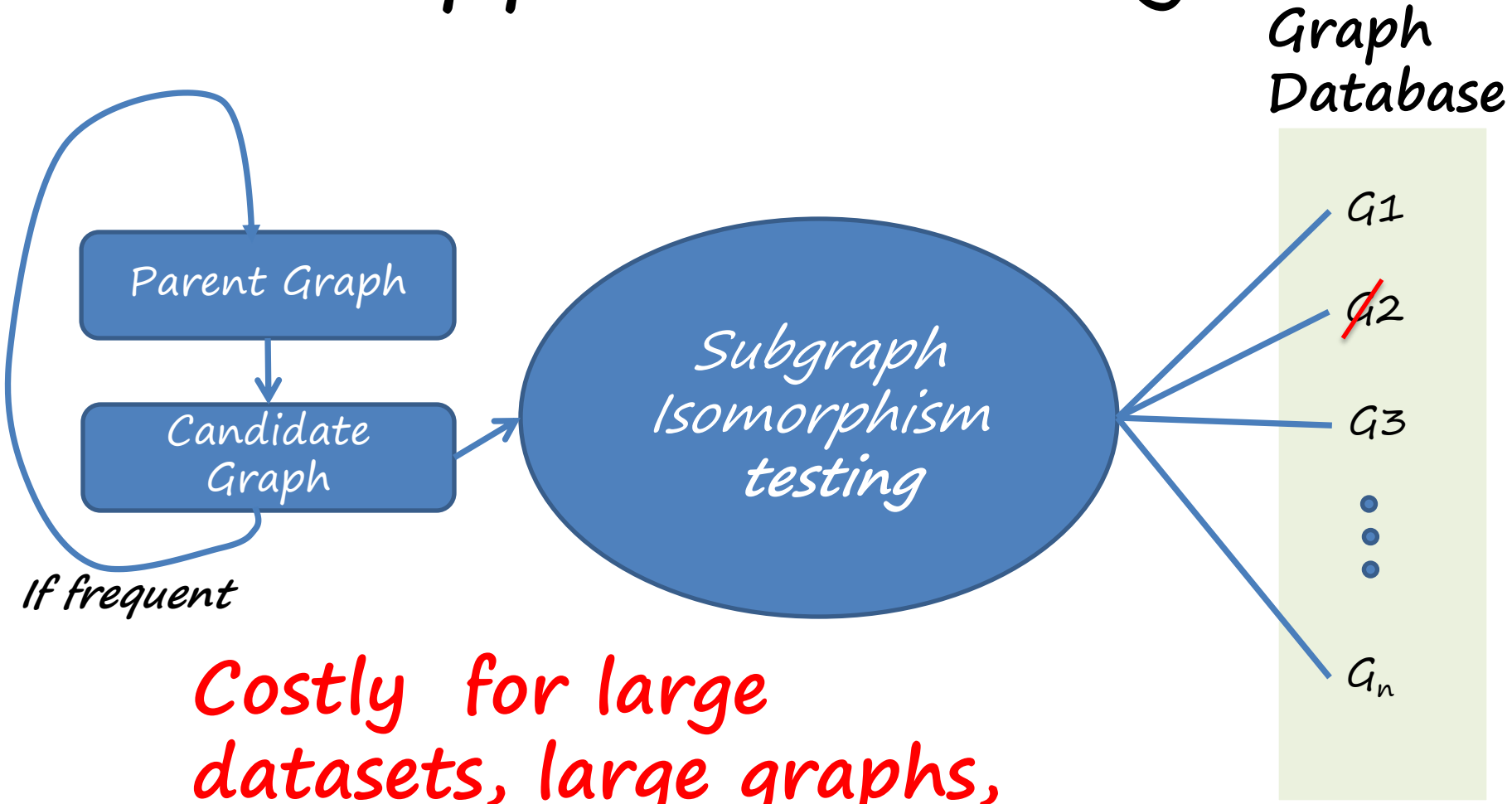


Frequent one-edge pattern, $\mathcal{F}_1$



Graph isomorphism

Support Counting

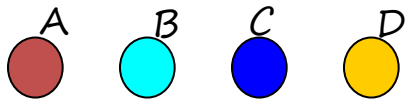


Costly for large datasets, large graphs, small support

Grand Unified Theory?

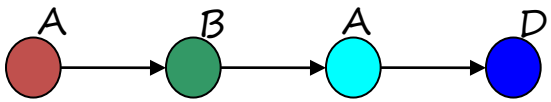
- Data Mining Template Library
  - Generic data structures & algorithms
  - Graphs, Trees, Sequences, Itemsets
  - Open-source; downloaded over 5300 times from dmtl.sourceforge.net
  - DMKD'08

SET



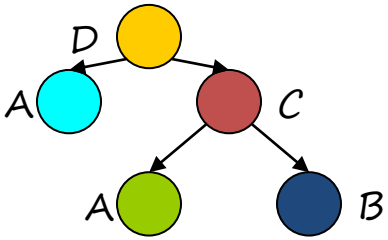
not connected
unique labels

SEQUENCE



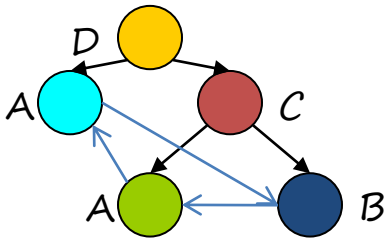
connected
directed
acyclic
in-degree <= 1
out-degree <=1

TREE



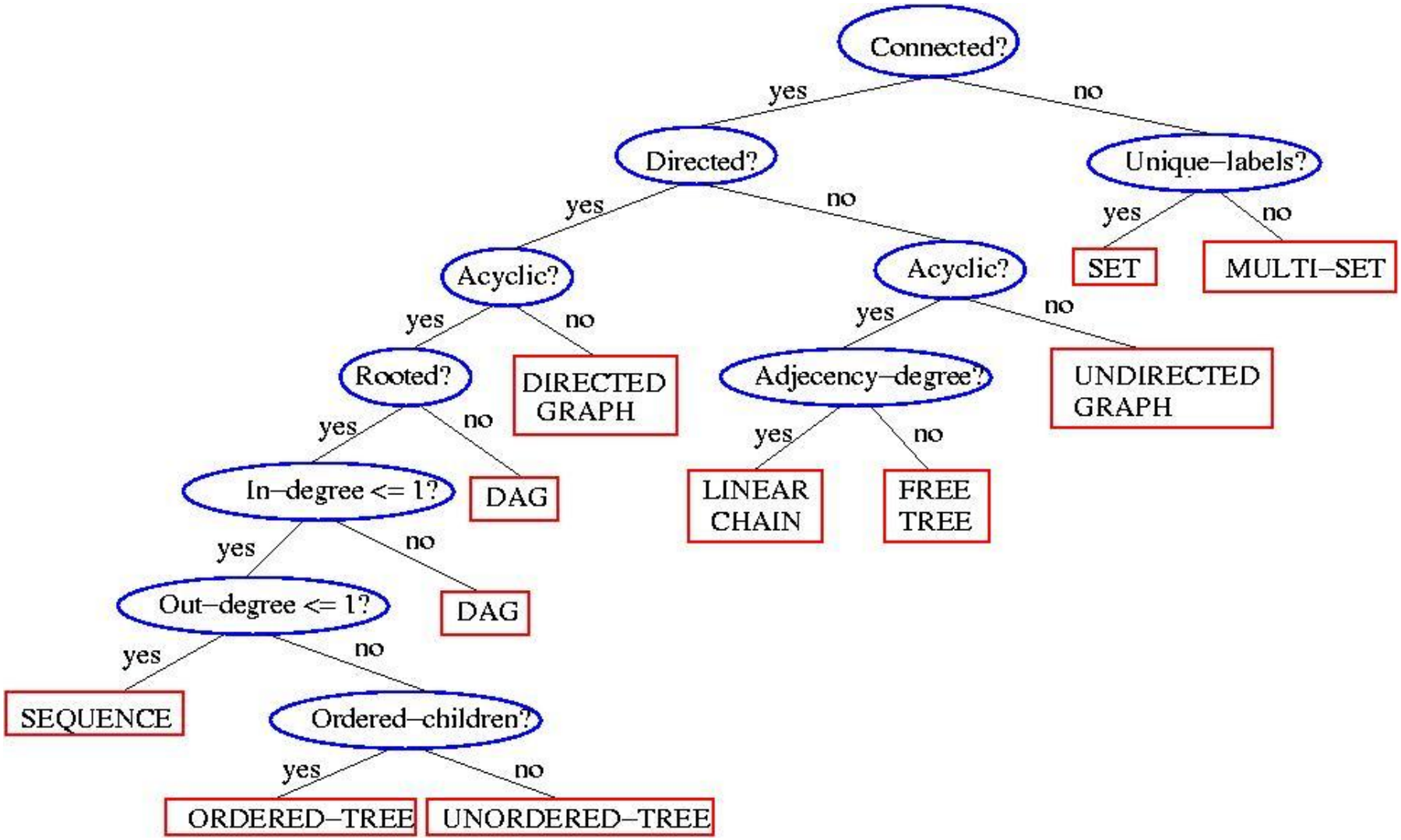
connected
directed
acyclic
in-degree <= 1
ordered

GRAPH



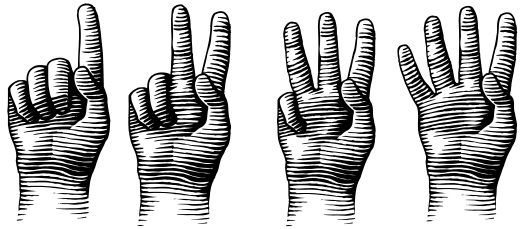
connected

Property Tree (Extensible)



What's good about frequent pattern mining?

- Fundamental exploratory mining task
- Very efficient algorithms for counting

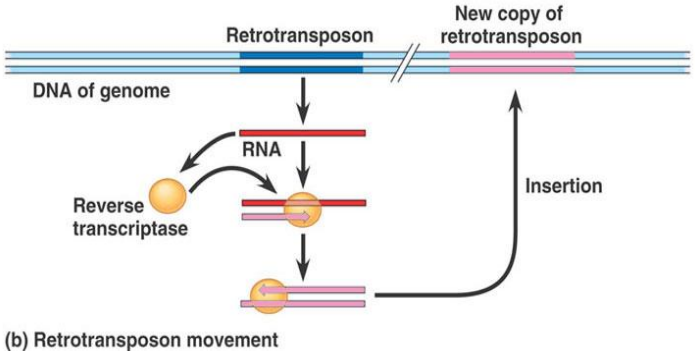


- Fast counting a basis for advanced statistical methods
  - Both Frequentist & Bayesian

- Patterns a basis for advanced kernel methods
- Varied applications: examples from bioinformatics

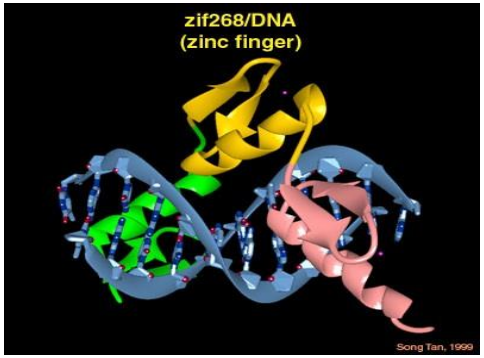
Structured Sequence Motifs

- Jumping Genes: LTR Retrotransposons
  - Ty1 Copia Motif in *A. thaliana*
  - TNGA[12,14]TWNYTNNA[19,21]
    TNTMYRT[4,6]WNCCNNNNRG
    [72,95]TGNNA[100,125]
    TNTANRTNRAYGA



(b) Retrotransposon movement

- Composite Regulatory Patterns
  (Transcription Factors)
  - UASH-URS1 cooperative factors in *Saccharomyces cerevisiae* (Yeast)
  - Involved in early meiotic expression during sporulation
    NNDTBNGDWGDNNDH[5,179]WBRGCSGCYVW

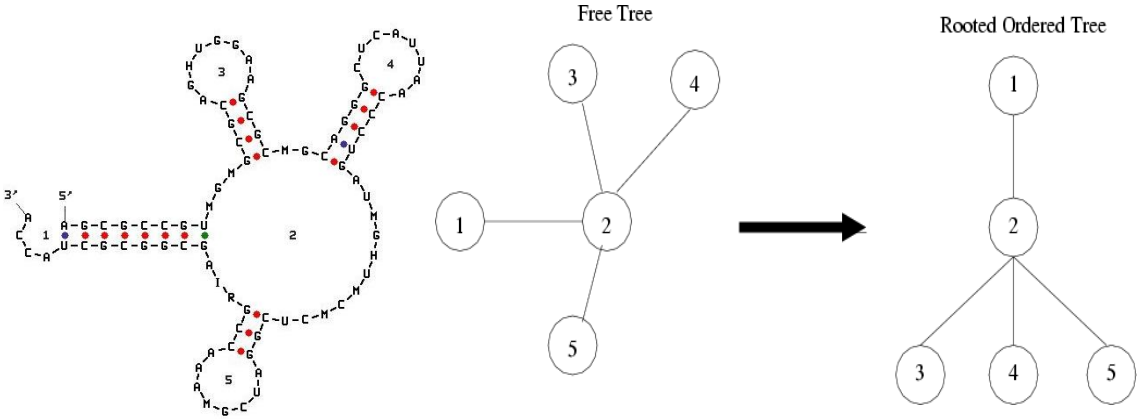


AMB'06

What does tree mining have to do with Mining Consensus RNA Motifs

- DB: 34 Eukarya RNA (RnaseP DB)
  - Ribonucleoprotein endonucleases that helps cleave transfer RNA precursors
  - Convert them into trees (RNA-as-Graph DB)
- Can also mine RNA foldings

RnaseP B. Subtilis

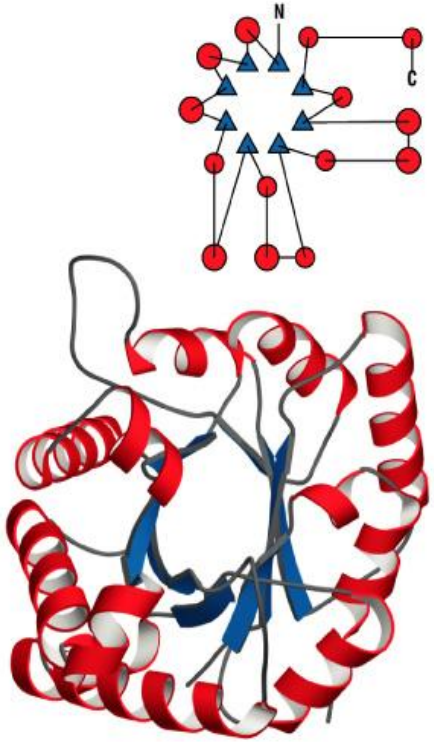


tRNA 2D structure converted to trees

TKDE'05

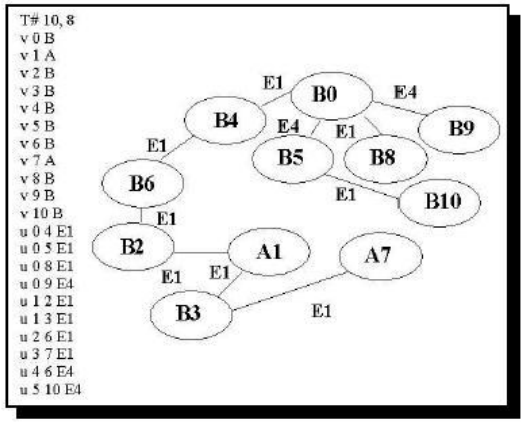
Here come the graphs: Protein Structural Motifs

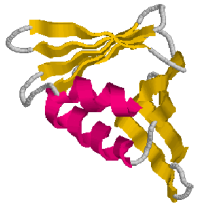
From **Protein Structure and Function** by Gregory A Petsko and Dagmar Ringe

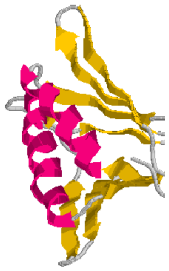


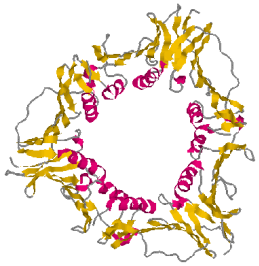
© 1999–2004 New Science Press

Mined Motif





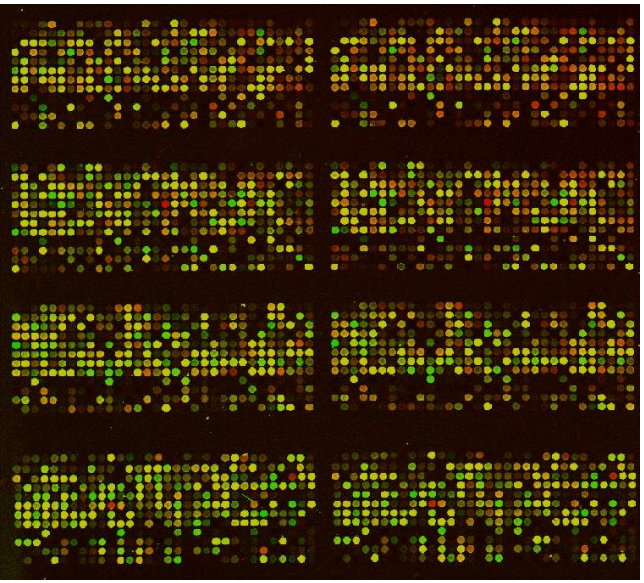




1CE8 1PLQ 1CZD

DNA Polymerase Factor Motif

Microarray Gene Expression Analysis: Coherent Clusters

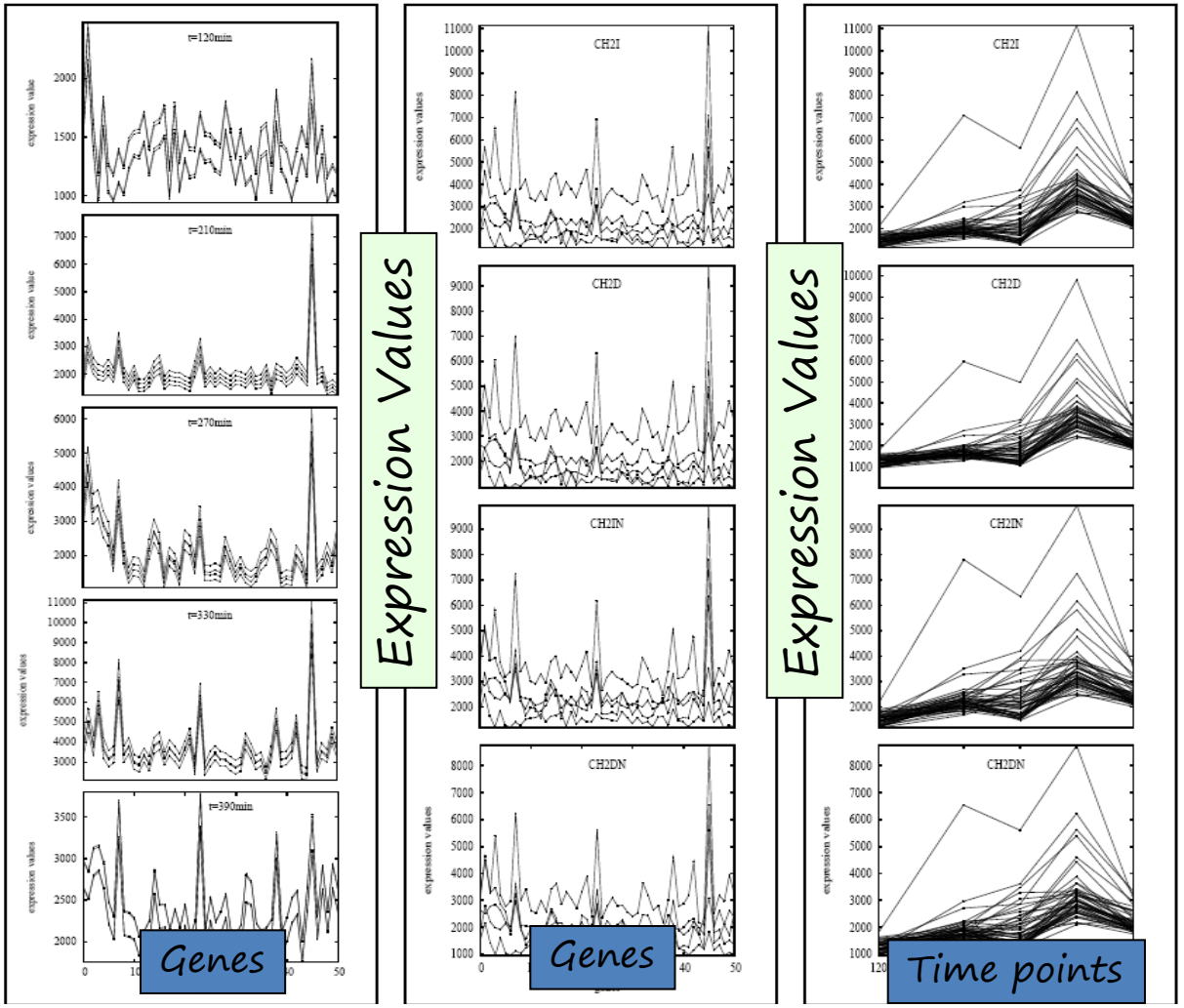


1	2	5
1	2	5
1	2	5

1	1	1
2	2	2
5	5	5

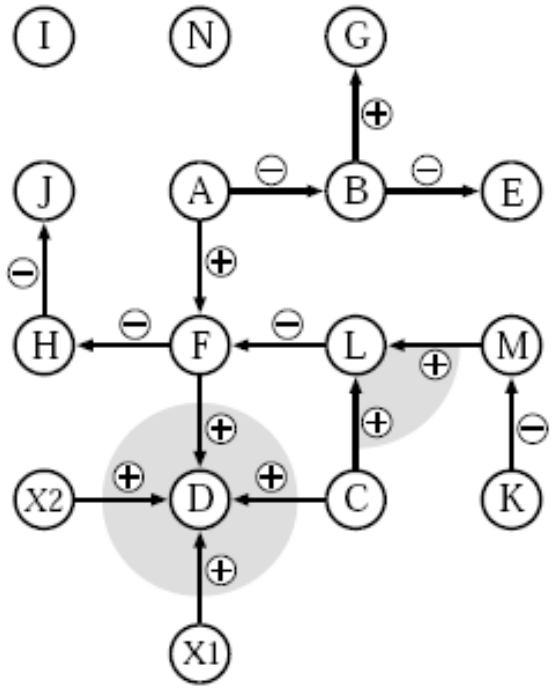
4	1	7
3	2	5
6	3	8

SIGMOD'05



Gene Networks: Boolean Expressions (AND, OR, CNF, DNF)

- Genes are involved in complex regulatory networks
- Can be represented as boolean networks
- Example: 16 genes
- ⊕: Activation, ⊖: Deactivation
- B, E, H, J, M are on if parents off
- G, L, D on if all parents on
  - D depends on C, F, X1, X2
- F on if A but not L
- A, C, I, K, N, X1, X2 don't depend on any other genes
- Generate a DB using 7 free genes: Truth table has 2^7=128 rows



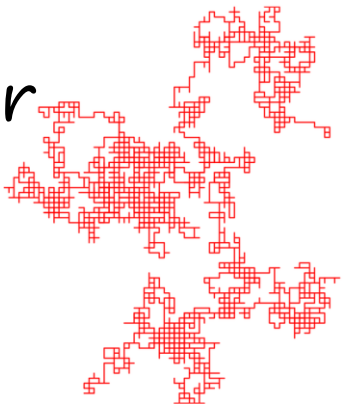
$$\left(\overline{D} \mid (A\,\overline{B}\,C\,E\,F\,\overline{G}\,\overline{H}\,J\,K\,\overline{L}\,\overline{M}\,X1\,X2)\right) \text{ AND}$$
$$\left(\overline{L} \mid (C\,\overline{F}\,H\,\overline{J}\,\overline{K}\,M)\right) \text{ AND}$$
$$\left((\overline{A}\,B\,\overline{E}\,G) \mid \overline{C} \mid D \mid L \mid \overline{X1} \mid \overline{X2}\right) \text{ AND}$$
$$\left((\overline{A}\,B\,\overline{E}\,G) \mid (C\,L) \mid (F\,\overline{H}\,J)\right) \text{ AND}$$
$$\left((\overline{F}\,H\,\overline{J}) \mid (A\,\overline{B}\,\overline{C}\,E\,\overline{G}) \mid (A\,\overline{B}\,E\,\overline{G}\,K\,\overline{M})\right)$$

SIGKDD'06

The death of complete pattern mining?

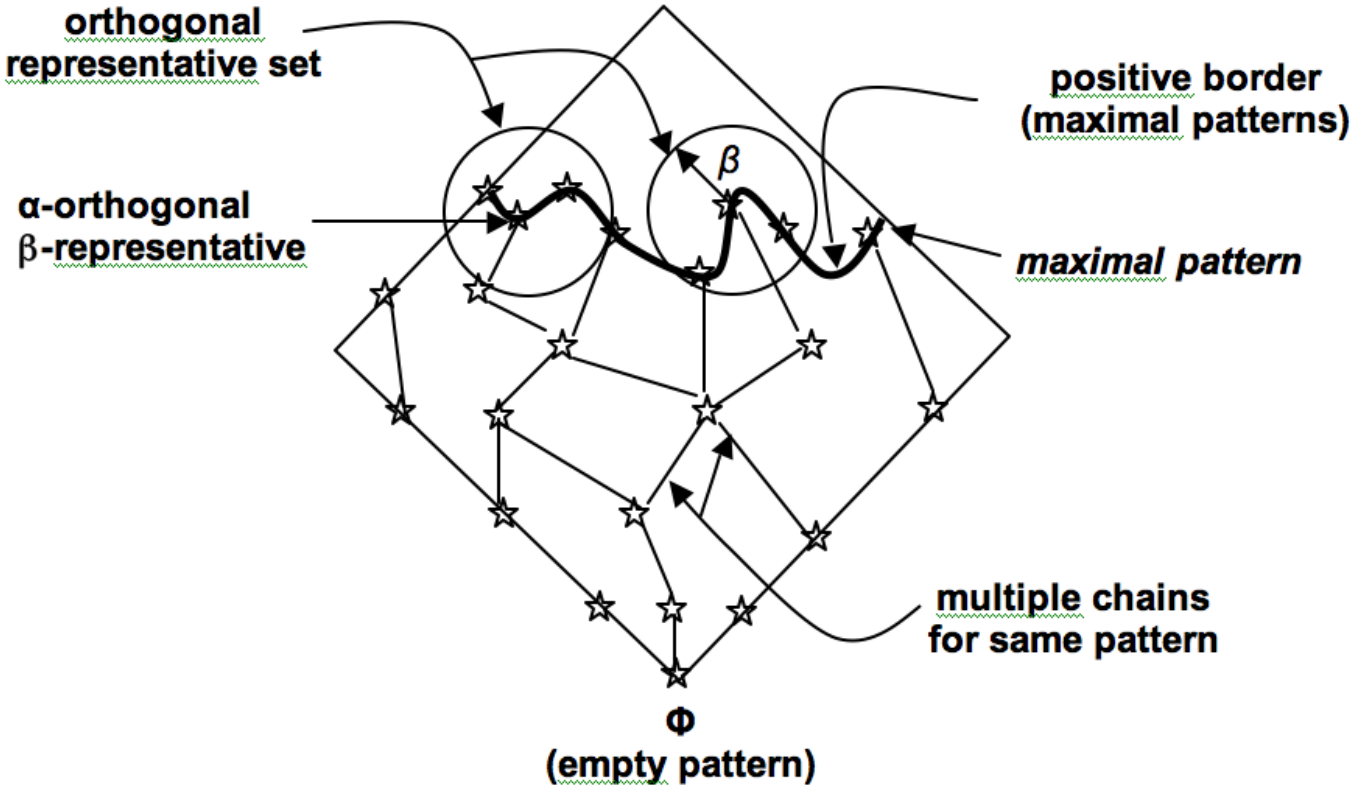
- The attack of the isomorphisms
- The era of complete enumeration of all frequent patterns over!
  - Infeasible in real world graphs
    - 3 graphs (genome-wide protein networks: pathways, gene expression, interactions), average 2154 nodes & 81607 edges (3MB total size)
    - Tried gSpan, Gaston, DMTL
    - Could not mine even at 100% support:

      7GB output, 8 million subgraphs. Abort!
- For many applications a representative or summary set is enough
- How to sample interesting patterns?
  - Take a (random) walk!



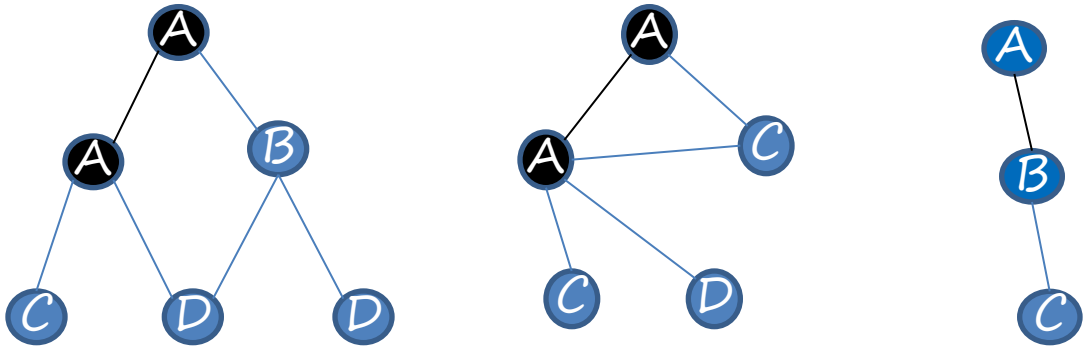
Sampling Maximal Subgraphs

- **random walks over chains** of subgraph partial order graph (POG): ORIGAMI (2007)

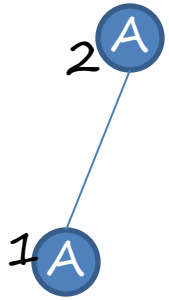


One Iteration: Walk over Chains

Graph DB

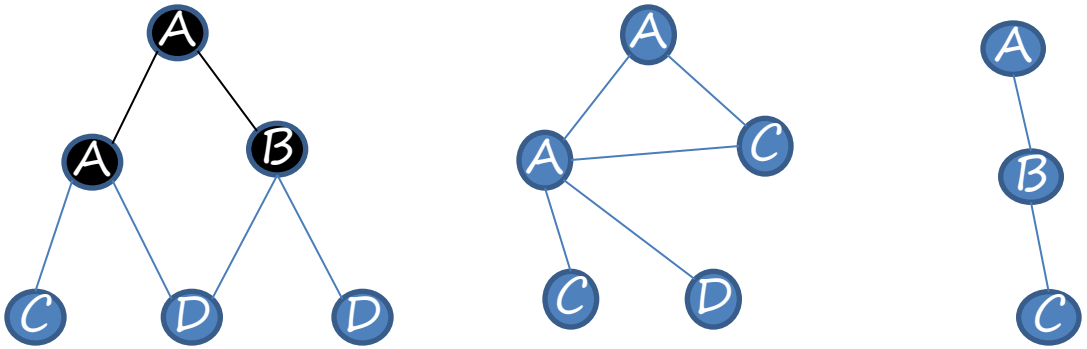


minsup =2

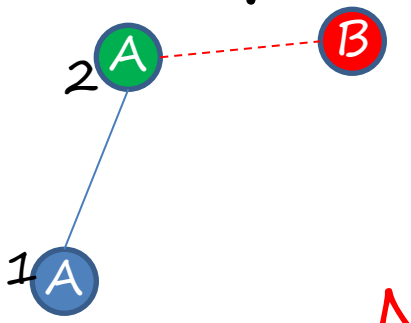


Pattern Extension …

Graph DB



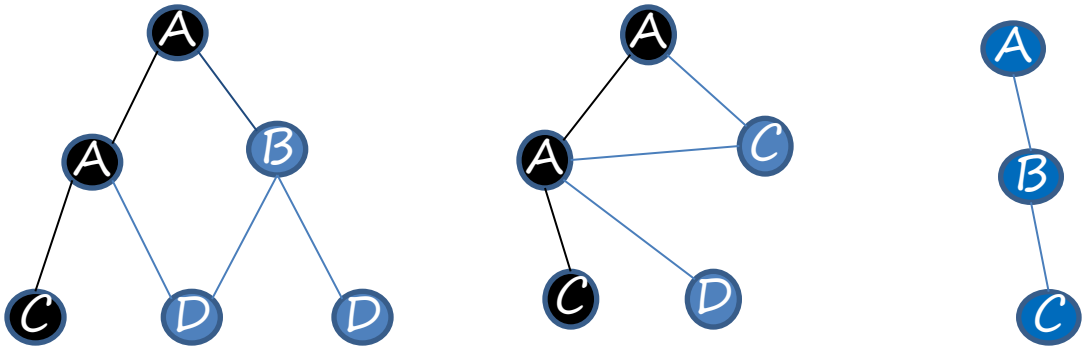
minsup =2



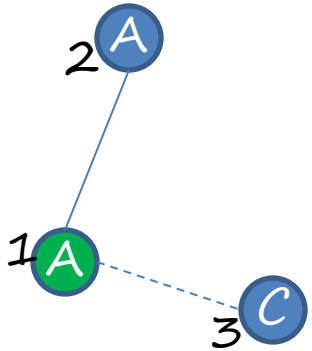
Not Frequent!

Pattern Extension …

Graph DB



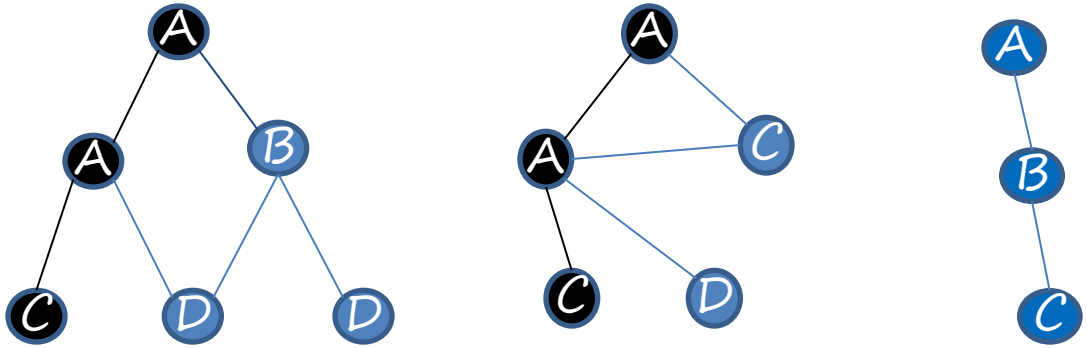
minsup =2



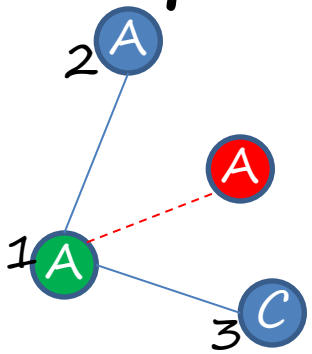
Frequent!

Pattern Extension …

Graph DB



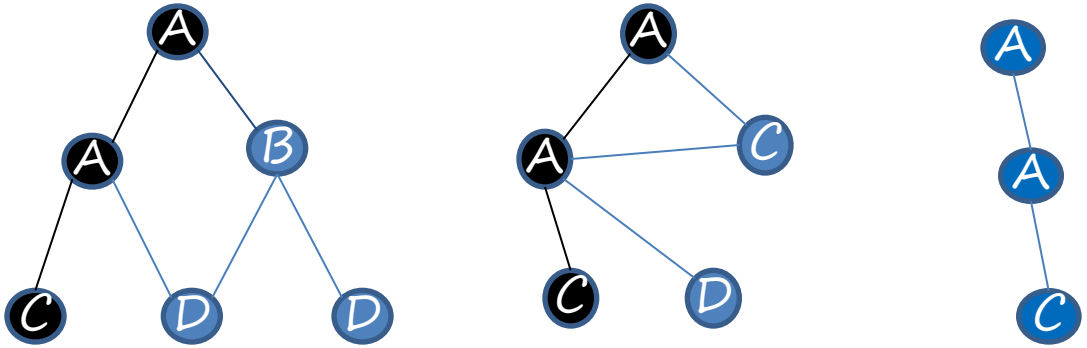
minsup =2



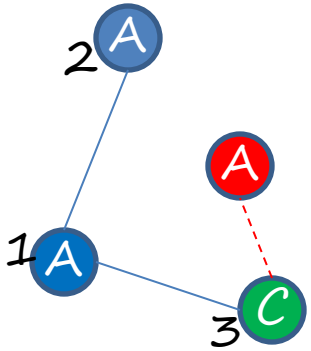
Not Frequent!
A—A has maximum frequency = 1

Pattern Extension …

Graph DB



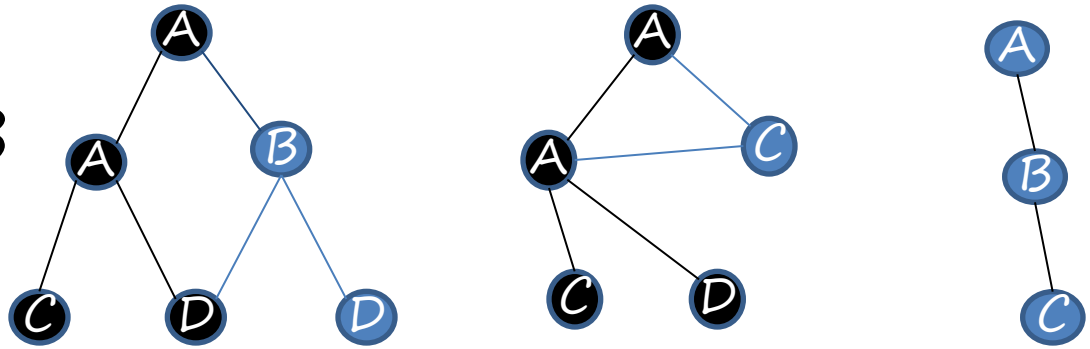
minsup =2



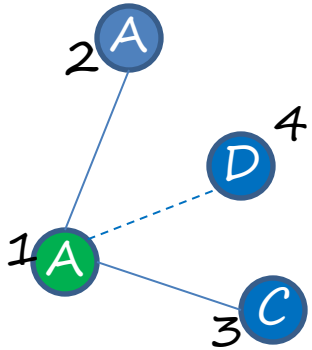
Not Frequent!

Pattern Extension …

Graph DB



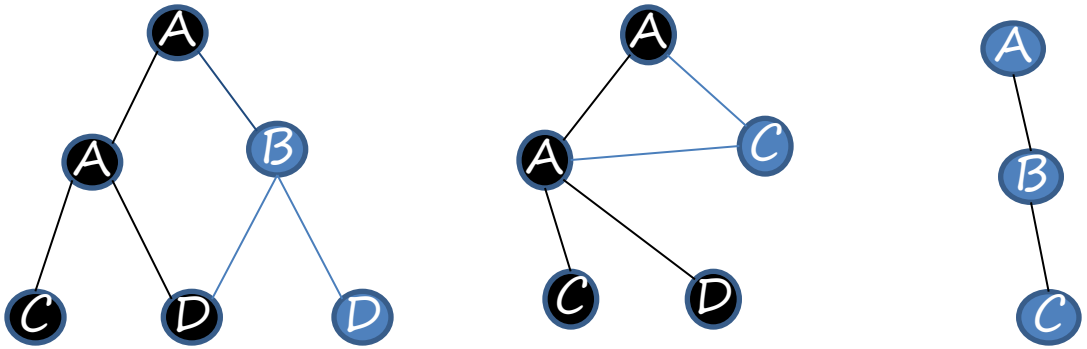
minsup =2



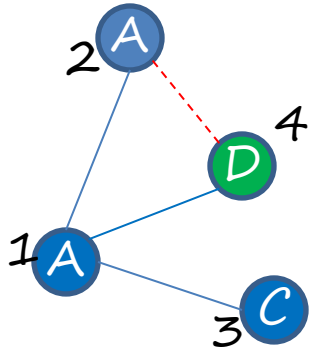
Frequent!

Pattern Extension …

Graph DB



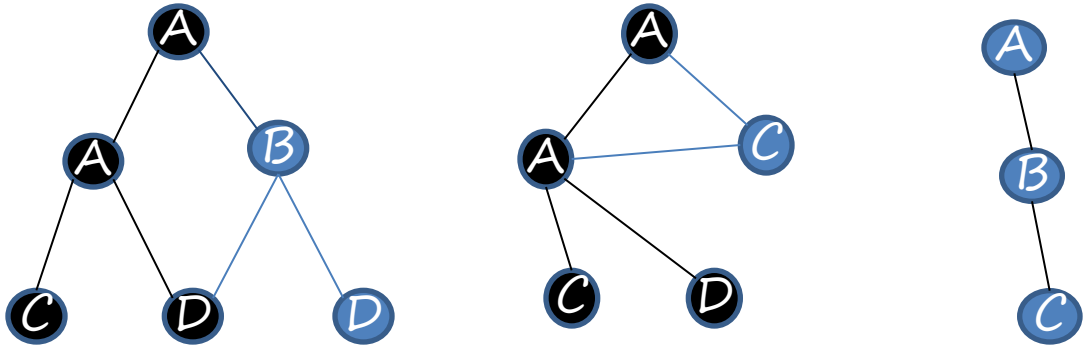
minsup =2



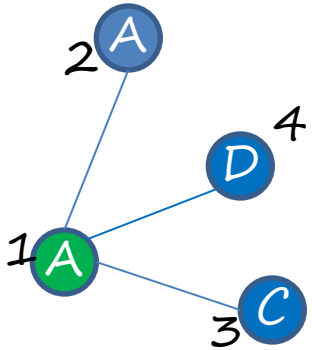
Not Frequent!
Edge A—D has
maximum frequency = 1

Are we there yet? random walk to maximality

Graph DB



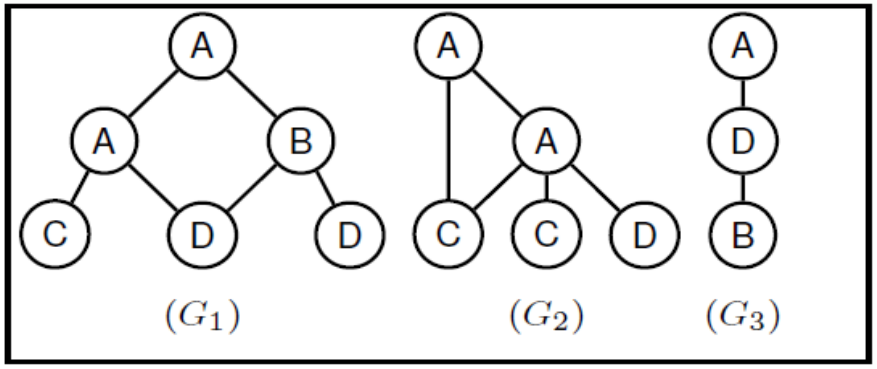
minsup =2



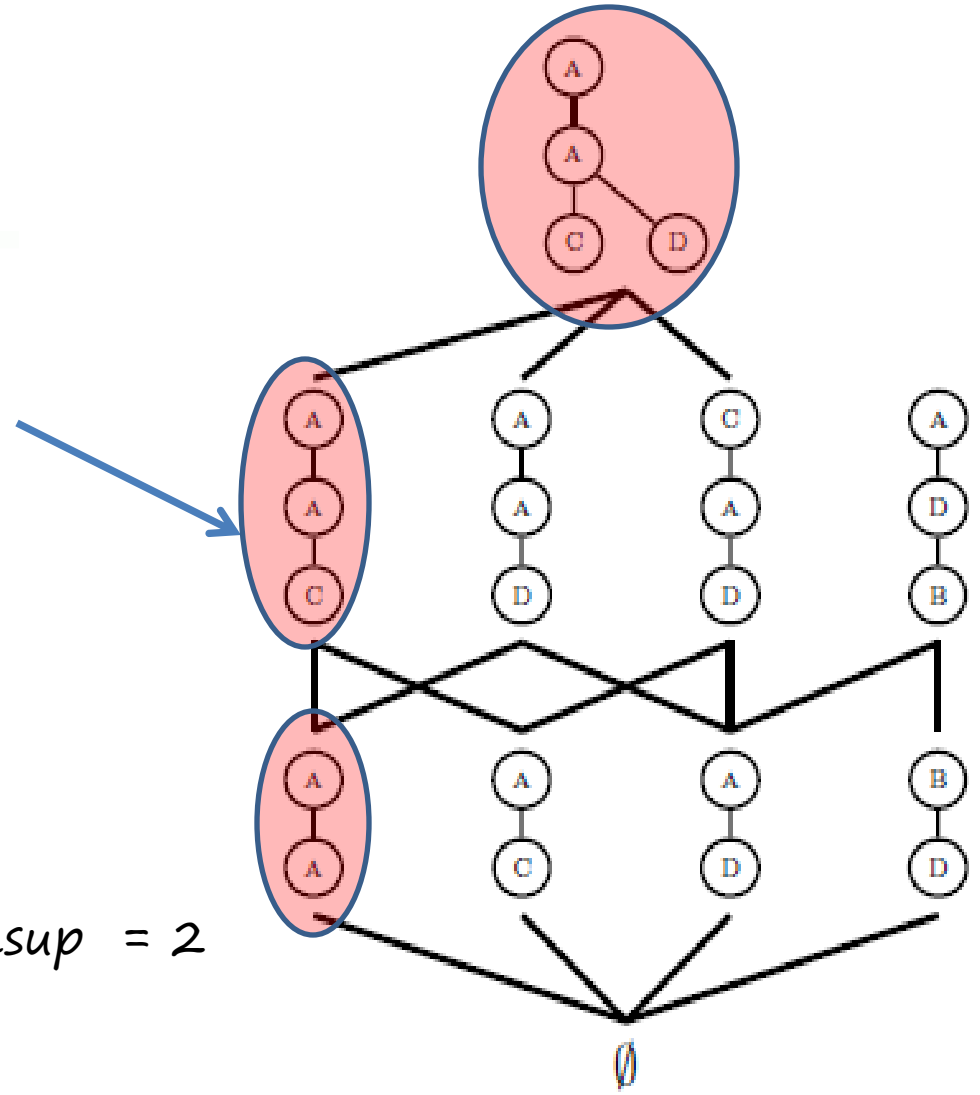
One iteration completed!

maximal frequent subgraph is obtained

Random Walk on POG Chains



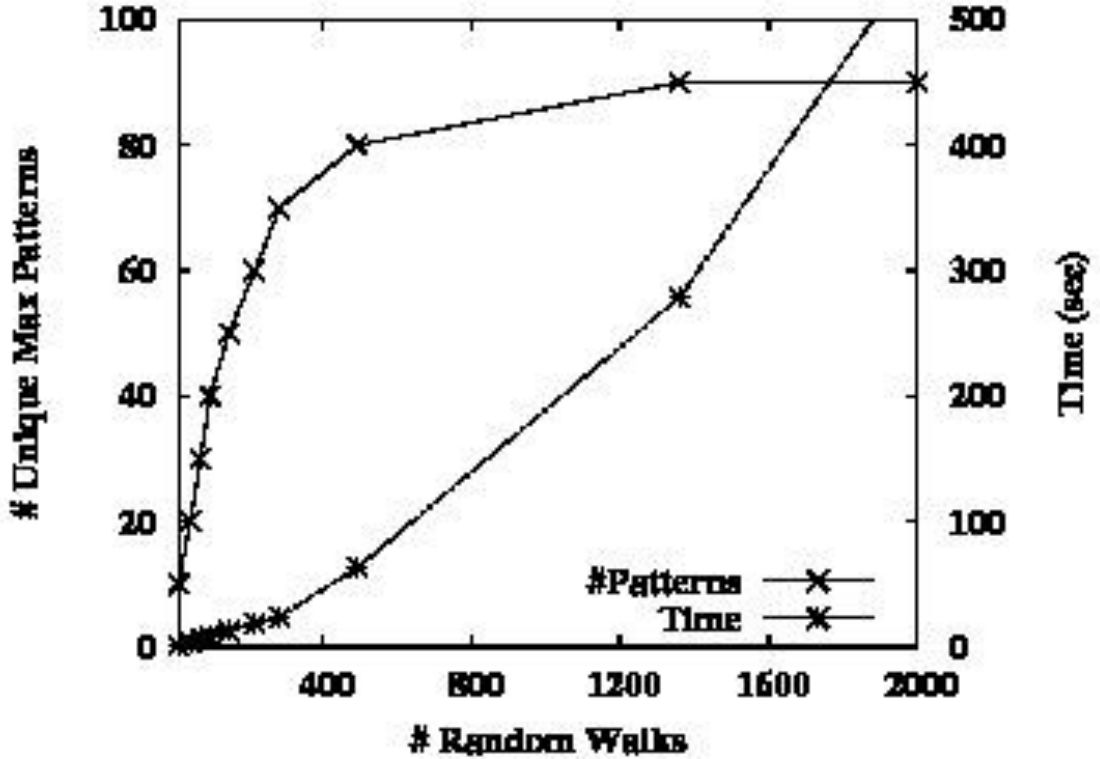
Graph Database, $\mathcal{D}$



POG Graph

Experimental Evidence

- All 300 maximal patterns found in 1400 iterations: total time 300 sec
- Complete methods terminated (7GB)



The good & bad

- ☺ Walks over chains is easy to implement
  - Minimal memory requirement
  - Each iteration yields one maximal pattern
  - Stop when k distinct patterns are mined
- ☹ No guarantee of uniform sample
  - If $e_1 e_2 \ldots e_m$ the sequence of random edge extensions, probability of the edge sequence

$$p(e_1, e_2, \ldots, e_m) = p(e_1) \prod_{i=2}^{m} p(e_i | e_1, \ldots, e_{i-1})$$

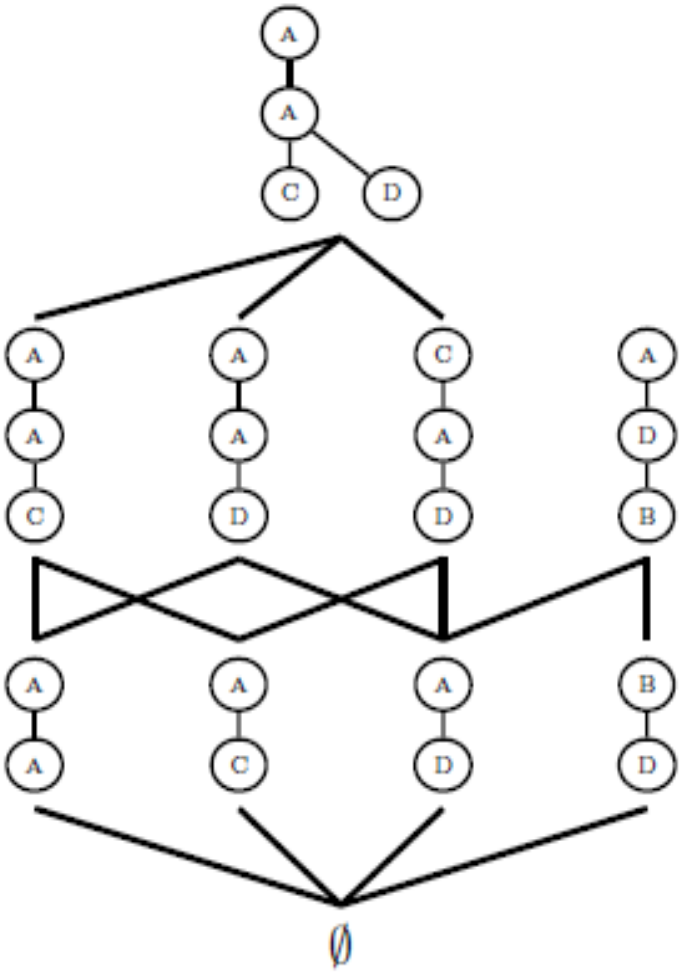
  - If a pattern has *ES* valid edge sequences, its generation probability is

$$\sum_{(e_1, e_2, \ldots, e_m) \in ES} p(e_1, e_2, \ldots, e_m)$$

  - Longer patterns have more valid paths, but probability is very small; small patterns prefered

Can uniformity be guaranteed? Markov Chain Monte Carlo Sampling

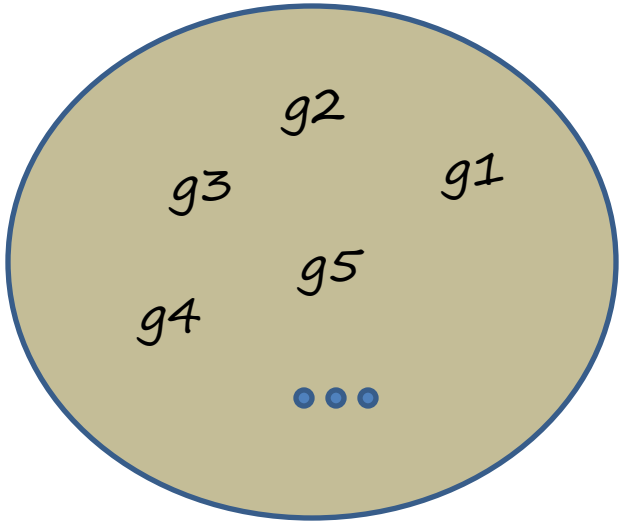
- POG as a transition graph
- Random walks on POG
- Local neighborhood
  - subgraph-supergraph
- Local transition probability



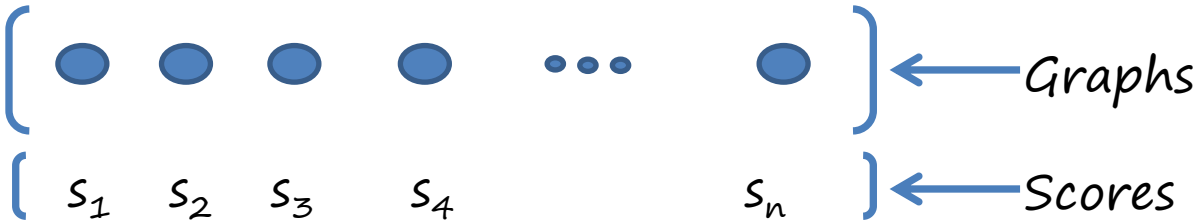
POG as transition graph

MCMC Challenges

- POG unknown
  - Don't want to know
- Complete statistics about frequent subgraphs unknown.
- Target distribution is not known entirely



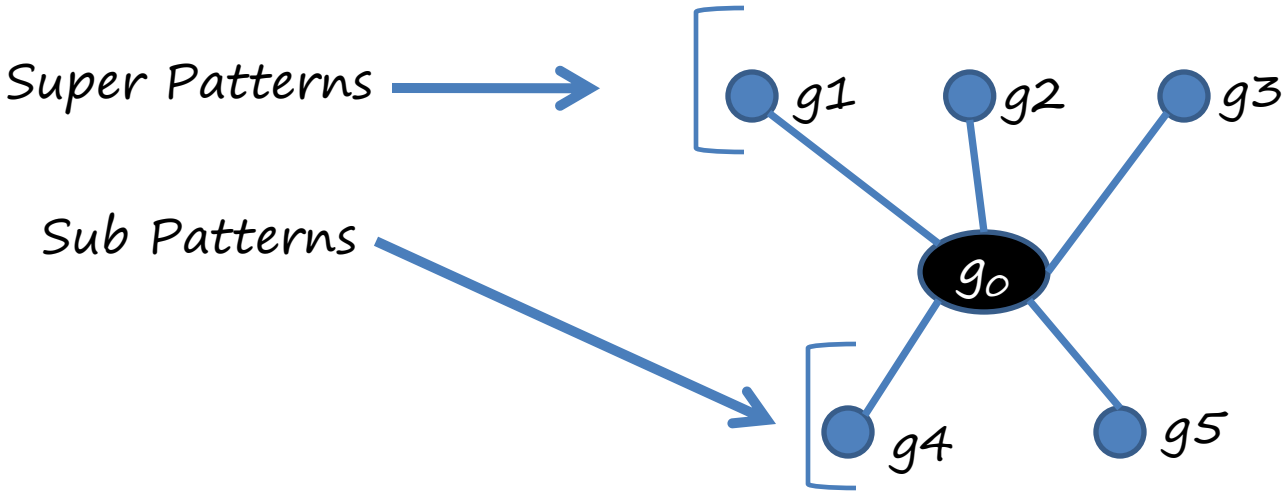
Output Space of Graph Mining: POG



We want,

$$\pi(i) = \frac{s_i}{\sum_i s_i}$$

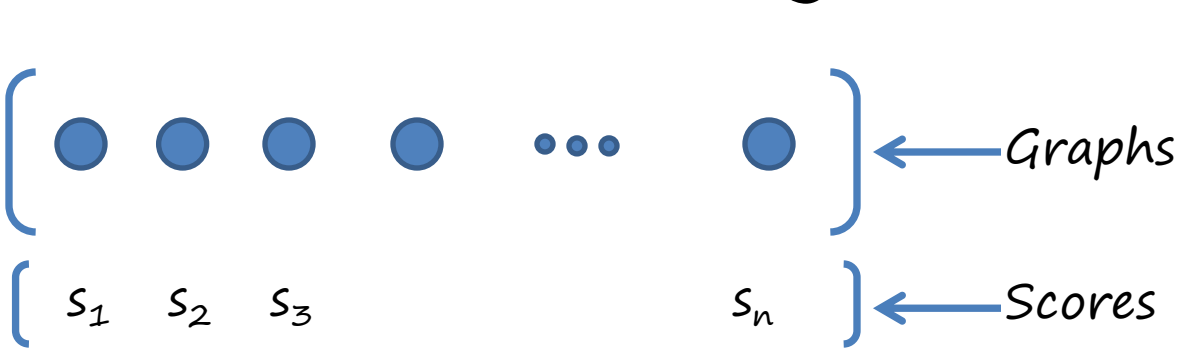
Local Computation of POG



Pattern that are not part of the output space is discarded during local neighborhood computation

$$\sum \begin{array}{|c|c|c|c|c|c|} \hline g1 & g2 & g3 & g4 & g5 & u \\ \hline p_{01} & p_{02} & p_{03} & p_{04} & p_{05} & p_{00} \\ \hline \end{array} = 1$$

Computing Transition Matrix P (achieve Target Distribution)



We want,

$$\pi(i) = \frac{s_i}{\sum_i s_i}$$

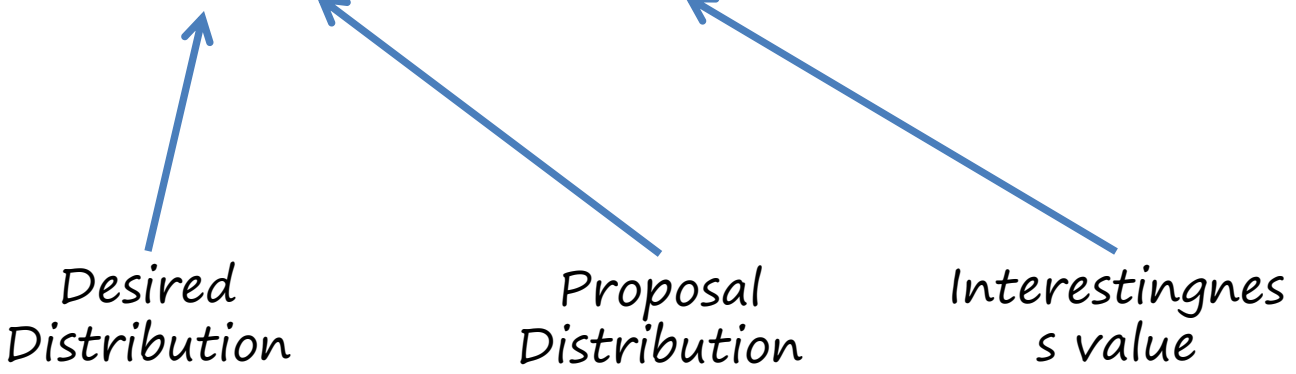
- If π is the stationary distribution, and P is the transition matrix, at equilibrium, we have,

$$\pi = \pi P$$

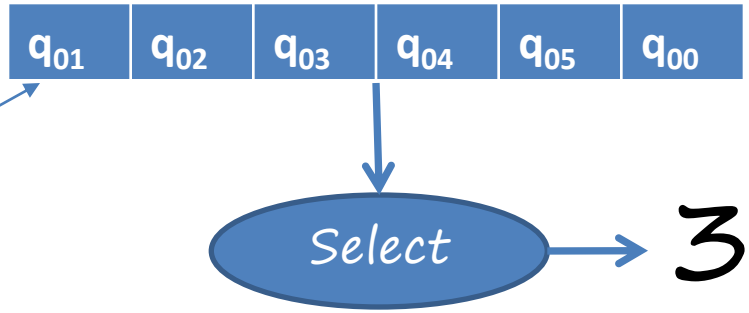
- Main task is to choose P, so that the desired stationary distribution is achieved
- Compute only one row of P (local computation)

Acceptance Probability Computation

$$\alpha_{ij} = \min\left\{\frac{\pi(j)}{\pi(i)}\frac{q_{ji}}{q_{ij}}, 1\right\} = \min\left\{\frac{b_j}{b_i}\frac{q_{ji}}{q_{ij}}, 1\right\}$$



Metropolis-Hastings Algorithm



1. Fix an arbitrary proposal distribution beforehand (q)

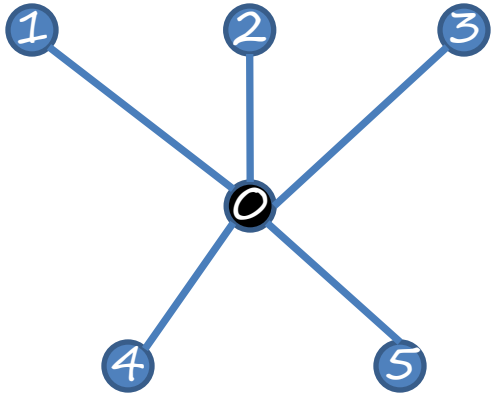
2. Find a neighbor j (to move to) by using the above distribution

$$\alpha_{03} = \min\left\{\frac{s_3 q_{30}}{s_0 q_{03}}, 1\right\}$$

3. Compute acceptance probability and accept the move with this probability

$$\alpha_{ij} = \min\left\{\frac{\pi(j)\ q_{ji}}{\pi(i)\ q_{ij}}, 1\right\} = \min\left\{\frac{b_j\ q_{ji}}{b_i\ q_{ij}}, 1\right\}$$

4. If accept move to j; otherwise, go to step 2

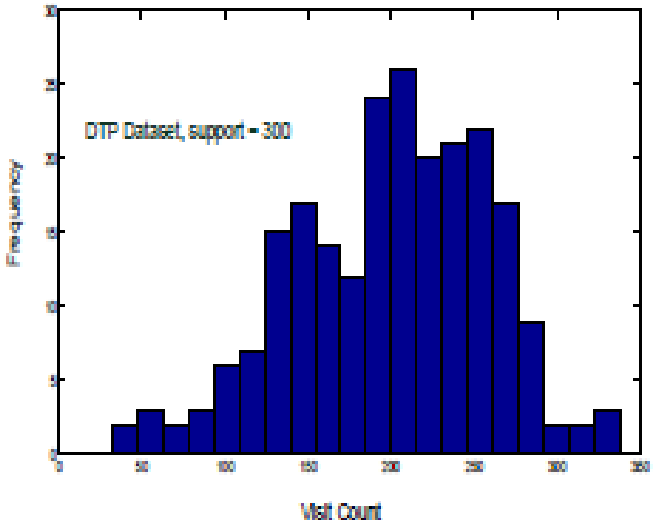


Different Sampling Tasks

- Uniform Sampling of Frequent Patterns
  - To explore the frequent patterns
  - To set a proper value of minimum support
  - To perform approximate counting
- Support Biased Sampling
  - To find Top-k Patterns in terms of support value
- Discriminatory subgraph sampling
  - Find subgraphs that are good features for classification
- Uniform Sampling of Maximal Pattern
  - For summarization of frequent patterns

Uniform Sampling of all Frequent Patterns

- Experiment Setup
  - Run the sampling algorithm for sufficient number of iterations and observe the visit count distribution
  - For a dataset with n frequent patterns, we perform 200*n iterations

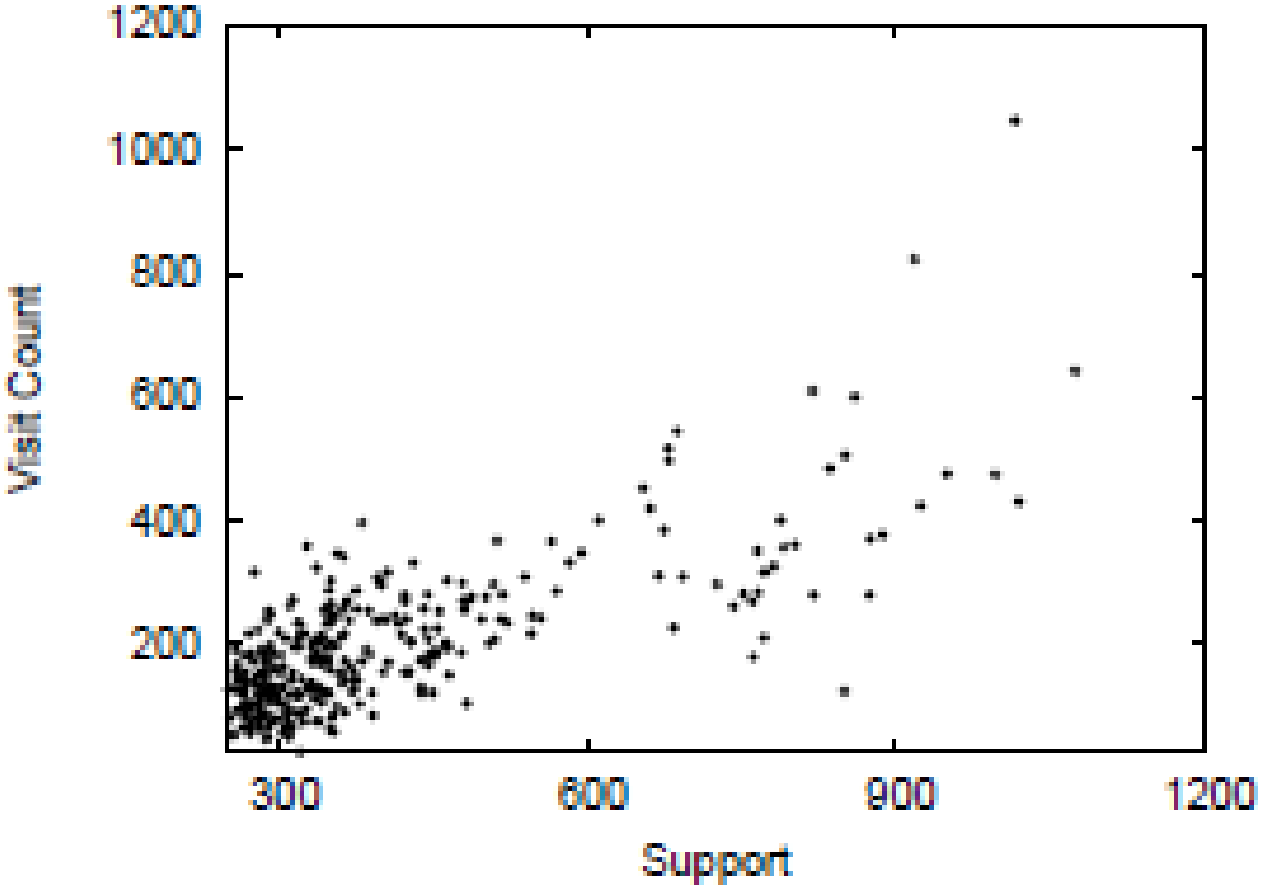


DTP Chemical Dataset
1084 graphs: 43/45 V/E

Uniform Sampling			
Max	Min	**Median**	**Std**
338	32	209	59.02

Ideal Sampling	
Median	Std
200	14.11

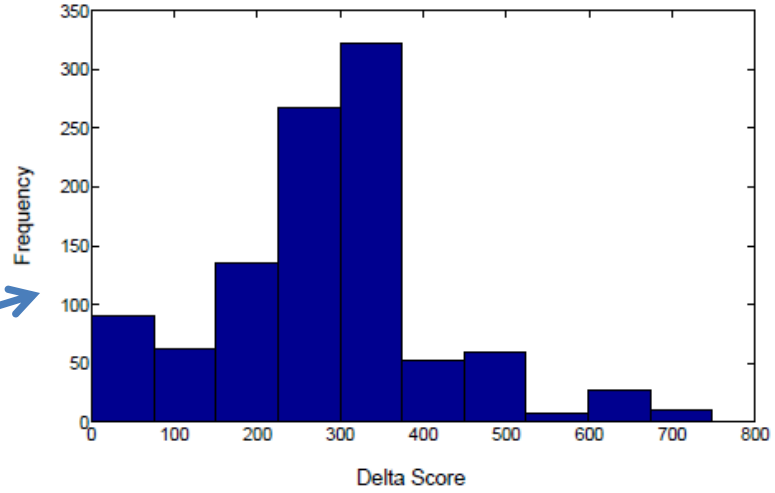
Support Biased Sampling



Scatter plot of Visit count and Support shows positive Correlation

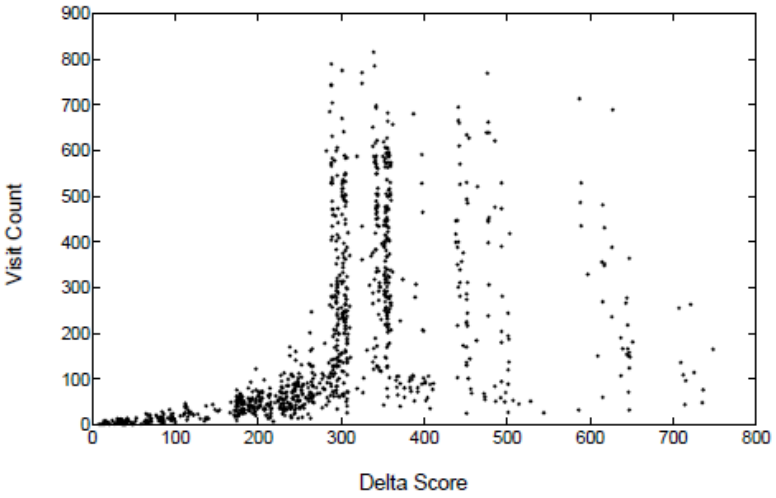
Discriminatory Sampling

Distribution of Discrimination Scores among all frequent Patterns

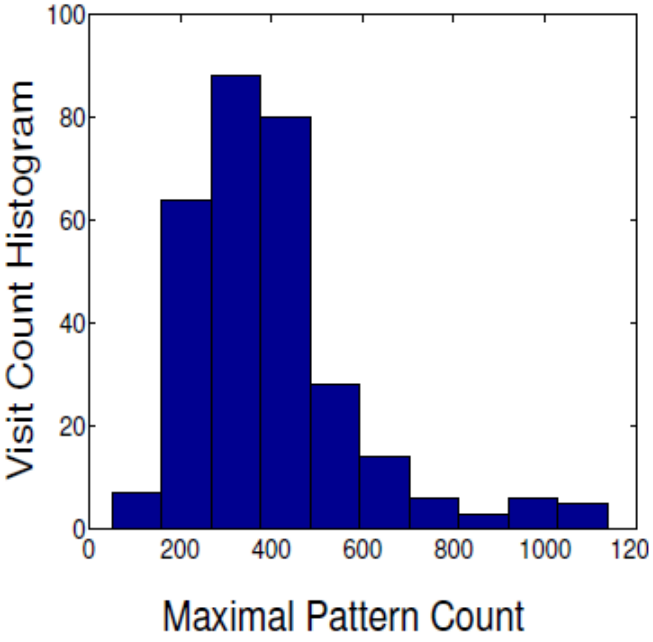


(a) Delta Score Distribution

Relation between Visit Counts and Discrimination Scores



Maximal Pattern Sampling



Our Algorithm		Ideal	
Median	Std	median	Std
367	185.7	400	19.97

Sampling Summary: The good & bad

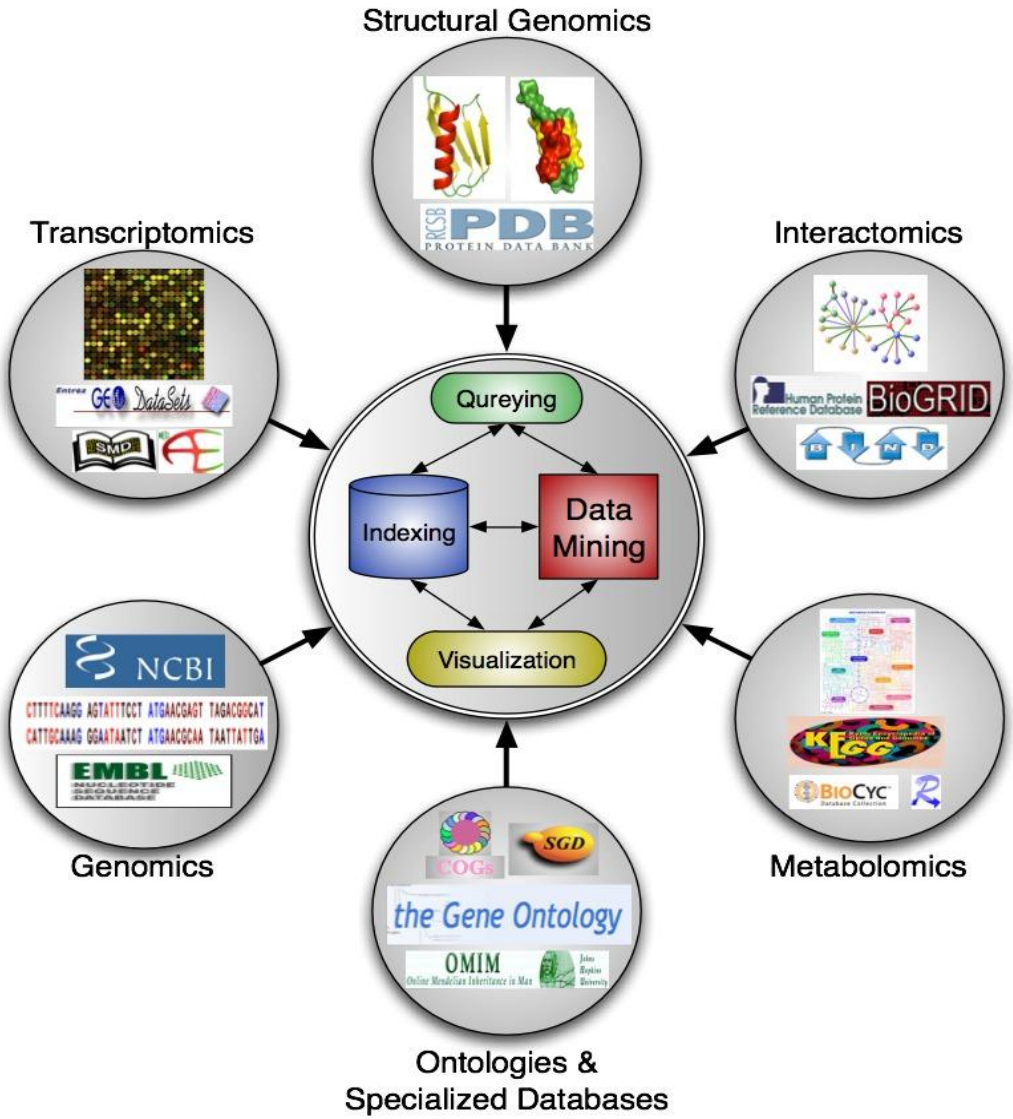
- ☺**Quality**: Sampling quality guaranty
- ☺**Scalability:** Visits only a small part of the search space
- ☺**Non-Redundant:** finds very dissimilar patterns by virtue of randomness
- ☺**Genericity:** In terms of pattern type and sampling objective

- ☹Efficiency still a concern for large graphs
  - support counting is still a bottleneck
  - How to improve on the isomorphism checking
  - How to effectively parallelize the support counting

Where are we headed? Into the mouth of the beast!

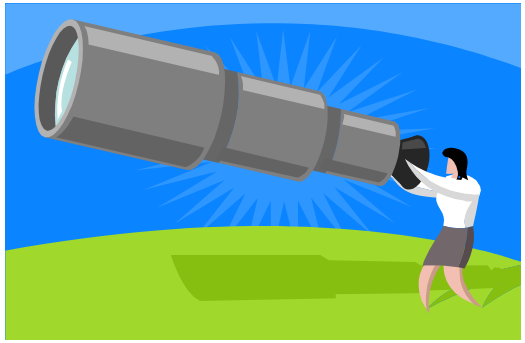
- Emergence of "complex" graphs
  - Enriched networks
    - Weighted
    - Multi-labeled (nodes & edges)
    - Temporal/spatial attributes
  - Distributed (multi-relational)
  - Uncertain
  - Dynamic
  - Massive & Unbounded (not known fully)
  - Networks, Networks & more Networks (everything is linked!)
    - E.g. Omics in Systems Biology, Semantic Web, Social Networks, …

Example: Mining the Omics Graph



Future of Pattern Mining

- Integrated Mining over enriched graphs and networks
- Constraints: Application oriented mining
- Approximate and uncertain pattern mining
- Dynamics & evolving pattern mining
- Sampling and summarization
- Patterns for Kernel Methods
  - Clustering (e.g., Spectral Methods)
  - Classification (e.g., Graph kernels)
- Grand Unified Theory Revisited
  - Bridge the gap between social network research, combinatorial pattern mining, bioinformatics, and data mining



The future's so bright, I gotta wear shades!